

**“Hydrological assessment of atmospheric forcing uncertainty in the Euro-Mediterranean area using a land surface model”
by Gelati et al.**

Cover letter to the editor

20 February 2018

Dear Dr. Matthias Bernhardt,

The authors' response to the comments of the two anonymous referees was published on the HESS web site.

All changes relative to the discussion paper are detailed in the marked-up version of the new manuscript. They include all the response elements given by the authors in response to the reviewers' comments (green and blue for Reviewer 1 and 2, respectively).

Yours sincerely,

Jean-Christophe Calvet, Emiliano Gelati.

LIST OF CHANGES MADE IN RESPONSE TO COMMENTS OF REVIEWER #1

1.3 [Page 5 – Line 22: please define the finer spatial scale.]

Response 1.3: changes in the marked-up version of the revised manuscript

P. 6, L. 20-20:

"In this study we use 12 columns within 0.5° latitude/longitude grid cells. The size of model grid cells is determined by the atmospheric forcing spatial resolution."

1.4 [Page 6 – Line 20: The adjustment of just one of the forcing variables (precipitation) leads to physical inconsistencies (Haddeland et al, 2012; Sippel et al, 2016). Authors could elaborate on this.]

Response 1.4: changes in the marked-up version of the revised manuscript

P. 17, L. 30-31:

"Limitations of (and alternatives to) bias correction methods adjusting a sub-set of the forcing variables or not involving statistical moments beyond the mean are discussed by Haddeland et al. (2012) and Sippel et al. (2016)."

1.5 [Page 7 – Line 10: Information regarding the impact of bias adjustment on large scale hydrological outputs are documented by Hagemann et al, 2011; Muerth et al, 2013; Papadimitriou et al, 2017.]

Response 1.5: changes in the marked-up version of the revised manuscript

P. 3, L. 20-22:

"AGCM-based atmospheric forcing may be bias-corrected to better reproduce observed climatology, thus introducing an additional layer of input uncertainty (Hagemann et al., 2011; Muerth et al., 2013; Papadimitriou et al., 2017)."

1.7 [Page 7 - Section: Atmospheric reference datasets. The atmospheric reference datasets that are used for comparison with the forcing are not, in some cases, independent. For example air temperature of WFDEI and PGF are bias corrected using datasets and compared against CRUv3.21. Several additional state of the art meteorological datasets exist and could be used, like for example:

- The Berkeley Earth Surface Temperatures (BEST) (Rohde et al, 2013)
- NASA Goddard's Global Surface Temperature Analysis (GISTEMP) (Hansen et al, 2010)
- Global Historical Climatology Network (Lawrimore et al, 2011)
- Global Soil Wetness Project dataset (GSWP3) (Yoshimura and Kanamitsu, 2013)

Authors could reflect on that.]

Response 1.7: changes in the marked-up version of the revised manuscript

P. 18, L. 1-10:

"Some of the used reference atmospheric datasets are not independent from the forcing. For example, CRU air temperature is used as reference and to bias-correct WFDEI and PGF forcing. The extensive use of CRU data is motivated by its relatively high update frequency and resolution (0.5°), which make this dataset attractive for land surface modelling. Moreover, since PGF and WFDEI are based on different reanalyses, their bias corrections yield different sub-diurnal variabilities. However, to reduce dataset interdependencies, future work on the evaluation of forcing uncertainties may benefit from using a wider ensemble of state of the art meteorological datasets, among others: the Global Soil Wetness Project (GSWP3) forcing dataset (Yoshimura and Kanamitsu, 2013); the Goddard Institute for Space Studies Temperature (GISSTEMP) analysis (Hansen et al., 2010); the Global Historical Climatology Network temperature dataset (Lawrimore et al, 2011); The Berkeley Earth Surface Temperatures (BEST) dataset (Rohde et al., 2013); and the Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset (Beck et al., 2017b)."

LIST OF CHANGES MADE IN RESPONSE TO COMMENTS OF REVIEWER #2

2.3 [The motivation of the study is to assess the modelling tool for planning human activities involving freshwater resources (i.e., integrated water resources management). I think this motivation is odd because normally, hydrologic models are used for this purpose instead of land surface models (LSMs). The former are designed for this purpose, they are more conceptual than the latter, and typically, parameters can be calibrated to achieve a satisfying representation of the terrestrial hydrologic cycle. On the contrary, LSMs are more physically-based, incorporate a wider range of processes (e.g., CO₂-cycle), and have been developed to provide the lower boundary condition for coupled atmosphere-land-ocean models over land.]

Response 2.3: changes in the marked-up version of the revised manuscript

P. 1, L. 10-11:

"Physically consistent descriptions of land surface hydrology are crucial for planning human activities that involve freshwater resources, especially in light of the expected climate change scenarios."

P. 2, L. 10 to P. 3, L. 13:

"Large-scale hydrology can be simulated using several approaches, ranging from lumped water balance models to distributed global hydrological models (GHMs) and land surface models (LSMs). LSMs and GHMs are used to study a wide range of water-related problems: hydrological and agricultural droughts (Dirmeyer et al., 2006; Szczypta et al., 2012; Szczypta et al., 2014); floods (Decharme et al., 2012; Pappenberger et al., 2012; Hirpa et al., 2016); agricultural production and irrigation (Rost et al., 2009; Jägermeyr et al., 2016); surface freshwater temperature and its impact on energy production (van Beek et al., 2012; van Vliet et al., 2012; Yearsley, 2012). LSMs, which were originally designed to provide lower boundary conditions and vertical fluxes for atmospheric global circulation models (AGCMs), generally simulate the diurnal course of energy exchanges, vegetation and carbon dynamics, and hydrology. GHMs, which were developed to estimate and compare freshwater availability with anthropic requirements, typically simulate processes that are relevant to water resources assessments such as, e.g., hydrodynamic routing, reservoir operation, and water demands (Bierkens, 2015). Recent model intercomparison exercises have evaluated the suitability of several GHMs and LSMs for simulating the continental water cycle globally (Haddeland et al., 2011; Schellekens et al., 2016; Beck et al., 2017a). While parameters of GHMs are generally calibrated using observed river discharge time series, those of LSMs are usually determined a priori from maps of land surface properties or expert judgement. Therefore, calibrated GHMs tend to reproduce river discharge better than uncalibrated LSMs (Beck et al., 2017a). However, models calibrated on river discharge are not guaranteed to perform better than uncalibrated models when considering other key hydrologic variables such as e.g. evapotranspiration (Schellekens et al., 2016). Since LSMs aim at reproducing several land surface fluxes and states, multi-criteria approaches may provide more robust parameter calibration and sensitivity analysis frameworks compared to traditional single-objective methods (Bastidas, 1998; Gupta et al., 1998; Gupta et al., 1999; Bastidas et al., 2006). Acknowledging that LSMs should be assessed using multiple criteria adds further complexity to the parameter calibration problem. Parameter calibration may cause model over-fitting, when the parameter set optimised over the calibration period proves to be sub-optimal in other periods (Andréassian et al., 2012). Over-fitting a physically based model may hinder the

detection of process misrepresentations and therefore model improvement. Over-fitting is of particular concern for climate change simulations, as models may be applied under climatic conditions that are very different from those of the calibration period (Knutti, 2008). Although model performance under climate change scenarios cannot be validated, uncalibrated models based on physically consistent process descriptions (such as LSMs) might be more robust than models based on parameterisations calibrated under past climate regimes and land uses (such as calibrated GHMs). For example, most GHMs do not simulate explicitly the surface energy fluxes that are crucial for evaporation, while most LSMs solve water and energy balances together (Haddeland et al., 2011; Pokhrel et al., 2012). Physically consistent descriptions of such fluxes are necessary for modelling land surface feedbacks into the atmosphere, which can be represented by coupling LSMs with AGCMs (Koster et al., 2004; Betts, 2007; Campoy et al., 2013). The outlined calibration trade-off should be carefully considered when choosing the modelling approach. In this study, we chose a LSM for assessing the hydrological impacts of atmospheric forcing uncertainties in the Euro-Mediterranean area. LSMs can be coupled to AGCMs (Widén-Nilsson et al., 2007) or run off-line, forced by gridded atmospheric datasets that are generally obtained from AGCM reanalysis or simulation."

2.4 [Nevertheless, I think that the evaluation of a LSM modelling system by different forcing data sets is a welcome contribution to the field of hydrology, but there are several criteria that have to be met to provide a meaningful analysis. The most important criteria is that the model satisfactorily reproduces the terrestrial hydrological cycle. Often, streamflow is used for this assessment and the authors also compare simulated streamflow against observations at 35 gauges (locations are shown in Figure 1). The median KGE over all gauges is at most 0.4 with at least 20\% of the gauges having a negative KGE, indicating a poor representation of the surface hydrology. This holds for all forcing data sets, which indicates that the poor performance is independent of these.]

Response 2.4: changes in the marked-up version of the revised manuscript

P. 12, L. 15-18:

"An aggregate score is often used as unique criterion in model calibration, because it allows applying efficient single objective optimisation algorithms. However, model calibration is an inherently multi-objective problem (Gupta et al., 1998). Thus, aggregate scores are not as informative as their individual components for evaluating model performance (Gupta et al., 2009)."

P. 15, L. 31 to P. 16, L. 2:

"While KGE does not provide diagnostic information on the causes of better or worse model performance, the individual scores show that forcing uncertainties have larger impacts on the mean and standard deviation of the simulated discharge than on shape, timing and inter-annual variability."

2.5 [As can be seen in the bottom row of Figure S1 in the supplements, most of the better performing gauges are nested sub-catchments of the Danube, Rhine, and Elbe river. These represent the same humid conditions and the good performance over the same area is "double counted". Other areas such as the cold region in North-Eastern Europe and catchments in the Mediterranean show significantly poorer performance.]

These also represent a large fraction of the European area. The North-Eastern part of Europe also experiences very poor agreement with soil moisture observations (Figure 3) and LAI (Figure 5). This poor agreement in this region does not allow any assessment of the forcing uncertainty because the model might lack important processes to reproduce the terrestrial hydrological cycle there. This might also be due to the fact that the authors use the parameters of previous work that has been only validated in the Rhone catchment (Decharme and Douville, 2006).]

Response 2.5: changes in the marked-up version of the revised manuscript

P. 13, L. 21-23:

"Possible causes for these low values may be: vegetation density, topography and soil frost occurrences, all of which are known to negatively affect SSM retrievals; and model misrepresentation of surface hydrological processes that are particularly relevant in cold and mountainous regions."

P. 18, L. 21-22:

"... investigating the improvement or inclusion in LSMs of physical processes that are relevant to topsoil hydrology at high latitudes, e.g. snowmelt, flooding and ponding (Gouttevin et al., 2013)."

2.7.1 [p. 20, l. 21 ff: The authors conclude that LSMs may progressively integrate groundwater modelling to improve the simulation of river discharge. This does not relate to the findings of this study. All the runs presented in this study include a linear reservoir to represent groundwater storage (see p. 5, l. 32f). The authors would have to present results that do not use this groundwater component to make this conclusion.]

Response 2.7.1: changes in the marked-up version of the revised manuscript

P. 20, L. 20-21:

"Vergnes and Decharme (2012) implemented a simple groundwater scheme coupled to the river channel in each model grid cell allowing bidirectional water exchanges through the riverbed, showing improvements in the simulations of total water storage variations and river discharge."

P. 22, L. 8-11:

"To improve the simulation of freshwater availability, LSMs may progressively integrate more physically based groundwater schemes (see e.g. Vergnes and Decharme, 2012; Vergnes et al., 2014) as well as lake and reservoir regulation (see e.g. Hanasaki et al., 10 2006; Pokhrel et al., 2012)."

Hydrological assessment of atmospheric forcing uncertainty in the Euro-Mediterranean area using a land surface model

Emiliano Gelati^{1,*}, Bertrand Decharme¹, Jean-Christophe Calvet¹, Marie Minvielle¹, Jan Polcher², David Fairbairn¹, Graham P. Weedon³

5 ¹CNRM, UMR 3589 (Météo-France, CNRS), Toulouse, France

²Laboratoire de Météorologie Dynamique du CNRS, UMR8539 (IPSL, CNRS), Paris, France

³Met Office, Joint Centre for Hydrometeorological Research, Wallingford, United Kingdom

*Now at: Joint Research Centre, European Commission, Ispra, Italy

Correspondence to: Jean-Christophe Calvet (jean-christophe.calvet@meteo.fr)

10 **Abstract.** Physically consistent descriptions of land surface hydrology are crucial for planning human activities that involve freshwater resources, especially in light of the expected climate change scenarios. We assess how atmospheric forcing data uncertainties affect land surface model (LSM) simulations by means of an extensive evaluation exercise using a number of state-of-the-art remote sensing and station-based datasets. For this purpose, we use the CO₂-responsive ISBA-A-gs LSM coupled the CNRM version of the Total Runoff Integrated Pathways (CTRIP) river routing model. We perform multi-forcing
15 simulations over the Euro-Mediterranean area (25°–75.5°N, 11.5°W–62.5°E, at 0.5° resolution) from 1979 to 2012. The model is forced using four atmospheric datasets. Three of them are based on the ERA-Interim reanalysis (ERA-I). The fourth dataset is independent from ERA-Interim: PGF, developed at Princeton University. The hydrological impacts of atmospheric forcing uncertainties are assessed by comparing simulated surface soil moisture (SSM), leaf area index (LAI) and river discharge against observation-based datasets: SSM from the European Space Agency's Water Cycle Multi-mission
20 Observation Strategy and Climate Change Initiative projects (ESA-CCI); LAI of the Global Inventory Modeling and Mapping Studies (GIMMS); and Global Runoff Data Centre (GRDC) river discharge. The atmospheric forcing data are also compared to reference datasets. Precipitation is the most uncertain forcing variable across datasets, while the most consistent are air temperature, and SW and LW radiation. At the monthly time scale, SSM and LAI simulations are relatively insensitive to forcing uncertainties. Some discrepancies with ESA-CCI appear to be forcing-independent and may be due to
25 different assumptions underlying the LSM and the remote sensing retrieval algorithm. All simulations overestimate average summer and early autumn LAI. Forcing uncertainty impacts on simulated river discharge are larger on mean values and standard deviations than on correlations with GRDC data. Anomaly correlation coefficients are not inferior to those computed from raw monthly discharge time series, indicating that the model reproduces inter-annual variability fairly well. However, simulated river discharge time series generally feature larger variability compared to measurements. They also
30 tend to overestimate winter-spring high flows and underestimate summer-autumn low flows. Considering that several differences emerge between simulations and reference data, which may not be completely explained by forcing uncertainty,

we suggest several research directions. These range from further investigating the discrepancies between LSMs and remote sensing retrievals to developing new model components to represent physical and anthropogenic processes.

1 Introduction

Freshwater resources are at the core of primary human needs. In particular, the production and supply of food and energy are closely interconnected with water availability and quality (Bazilian et al., 2011; Ringler et al., 2013; Lawford et al., 2013; Damerau et al., 2016). Climate change may add further constraints to the sustainable use of water resources, by causing increased drought frequency and intensity in vulnerable regions such as the Euro-Mediterranean area (Planton et al., 2012; IPCC, 2014). Therefore, understanding and predicting water cycle processes on continental surfaces are necessary to build planning and assessment tools for integrated water resources management.

Large-scale hydrology can be simulated using several approaches, ranging from lumped water balance models to distributed global hydrological models (GHMs) and land surface models (LSMs). LSMs and GHMs are used to study a wide range of water-related problems: hydrological and agricultural droughts (Dirmeyer et al., 2006; Szczypta et al., 2012; Szczypta et al., 2014); floods (Decharme et al., 2012; Pappenberger et al., 2012; Hirpa et al., 2016); agricultural production and irrigation (Rost et al., 2009; Jägermeyr et al., 2016); surface freshwater temperature and its impact on energy production (van Beek et al., 2012; van Vliet et al., 2012; Yearsley, 2012).

LSMs, which were originally designed to provide lower boundary conditions and vertical fluxes for atmospheric global circulation models (AGCMs), generally simulate the diurnal course of energy exchanges, vegetation and carbon dynamics, and hydrology. GHMs, which were developed to estimate and compare freshwater availability with anthropic requirements, typically simulate processes that are relevant to water resources assessments such as, e.g., hydrodynamic routing, reservoir operation, and water demands (Bierkens, 2015). Recent model intercomparison exercises have evaluated the suitability of several GHMs and LSMs for simulating the continental water cycle globally (Haddeland et al., 2011; Schellekens et al., 2016; Beck et al., 2017a). While parameters of GHMs are generally calibrated using observed river discharge time series, those of LSMs are usually determined a priori from maps of land surface properties or expert judgement. Therefore, calibrated GHMs tend to reproduce river discharge better than uncalibrated LSMs (Beck et al., 2017a). However, models calibrated on river discharge are not guaranteed to perform better than uncalibrated models when considering other key hydrologic variables such as e.g. evapotranspiration (Schellekens et al., 2016). Since LSMs aim at reproducing several land surface fluxes and states, multi-criteria approaches may provide more robust parameter calibration and sensitivity analysis frameworks compared to traditional single-objective methods (Bastidas, 1998; Gupta et al., 1998; Gupta et al., 1999; Bastidas et al., 2006). Acknowledging that LSMs should be assessed using multiple criteria adds further complexity to the parameter calibration problem.

Parameter calibration may cause model over-fitting, when the parameter set optimised over the calibration period proves to be sub-optimal in other periods (Andréassian et al., 2012). Over-fitting a physically based model may hinder the detection of

process misrepresentations and therefore model improvement. Over-fitting is of particular concern for climate change simulations, as models may be applied under climatic conditions that are very different from those of the calibration period (Knutti, 2008). Although model performance under climate change scenarios cannot be validated, uncalibrated models based on physically consistent process descriptions (such as LSMs) might be more robust than models based on parameterisations calibrated under past climate regimes and land uses (such as calibrated GHMs). For example, most GHMs do not simulate explicitly the surface energy fluxes that are crucial for evaporation, while most LSMs solve water and energy balances together (Haddeland et al., 2011; Pokhrel et al., 2012). Physically consistent descriptions of such fluxes are necessary for modelling land surface feedbacks into the atmosphere, which can be represented by coupling LSMs with AGCMs (Koster et al., 2004; Betts, 2007; Campoy et al., 2013).

The outlined calibration trade-off outlined should be carefully considered when choosing the modelling approach. In this study, we chose a LSM for assessing the hydrological impacts of atmospheric forcing uncertainties in the Euro-Mediterranean area. LSMs can be coupled to AGCMs (Widén-Nilsson et al., 2007) or run off-line, forced by gridded atmospheric datasets that are generally obtained from AGCM reanalysis or simulation.

The lack of direct observations suitable for large-scale quantification of spatially heterogeneous hydrological variables (e.g. soil moisture and evapotranspiration) gives LSMs a key role in water resources assessments. However, to obtain reliable hydrological simulations from LSMs, several uncertainties need to be tackled. Their main sources are: model physics (Dirmeyer, 2011; Beck et al., 2017a); land surface parameters (Douville, 1998; Milly and Shmakin 2002); and atmospheric forcing (Ngo-Duc et al., 2005; Decharme and Douville, 2006b; Nasonova et al., 2011). The experiments of Guo et al. (2006b) found that atmospheric forcing uncertainties affect LSM hydrological simulations as much as uncertainties stemming from the models themselves. AGCM-based atmospheric forcing may be bias-corrected to better reproduce observed climatology, thus introducing an additional layer of input uncertainty (Hagemann et al., 2011; Muerth et al., 2013; Papadimitriou et al., 2017).

Uncertainty impacts can be assessed by comparing simulation outputs with available independent datasets of land surface state variables and fluxes. Such datasets should be representative at the current spatial scales of LSMs (10^1 – 10^2 km).

River discharge is useful for validating large-scale LSM simulations as it is the integral of all hydrological processes occurring in a catchment (Fekete et al., 2012). Moreover, observed river discharge time series are globally available for a large number of catchments (Hannah et al., 2011). LSMs can be coupled with river routing models that simulate horizontal channel water movement from headwaters to oceans (Ducharne et al., 2003; Pappenberger et al., 2010; Decharne et al., 2010). Anthropogenic alterations of the natural hydrological regime (e.g. reservoir regulation, abstractions, transfers and man-made drainage systems) pose challenges for achieving realistic LSM river discharge simulations (Li et al., 2013). While allowing comparisons with the outcome of simulated hydrological processes over a catchment, river discharge does not provide spatially distributed information about model performance at scales finer than the catchment size.

Remote sensing retrievals of the land surface constitute the ideal benchmarks for spatially distributed evaluations of LSMs (Overgaard et al., 2006), as they provide information integrated over areas whose size is compatible with model resolution.

Moreover, there is a growing availability of remote sensing products that are highly relevant for monitoring land-atmosphere interactions. Among those, retrievals of surface soil moisture (SSM) and leaf area index (LAI), which is the total photosynthetically active leaf area per ground unit area, are of particular importance. SSM and LAI are primary controls on the energy, water and carbon fluxes at the land-atmosphere interface. They are both key factors in determining evapotranspiration and surface albedo. SSM influences infiltration and surface runoff generation, providing the upper boundary conditions for moisture redistribution in unsaturated soils. LAI controls plant phenological development and canopy interception of precipitation.

Remote sensing estimates are affected by uncertainties that may stem from instruments, retrieval algorithms, or external atmospheric and land surface data necessary for the indirect estimation of geophysical variables. Thus, the cross-comparisons between remote sensing retrievals and LSM simulations provide useful insights for improving both products (see e.g. Brut et al., 2009; Lafont et al., 2012; Albergel et al., 2013a,b; Szczypta et al., 2014; Polcher et al., 2016; Barella-Ortiz et al., 2017). Furthermore, these exercises are useful for the development of land data assimilation systems that aim at integrating independent measurements of key state variables such as SSM and LAI into LSMs (Rodell et al., 2004; Draper et al., 2012; Barbu et al., 2014; Albergel et al., 2017).

Here we present a hydrological evaluation of the impacts of atmospheric forcing uncertainties on LSM simulations. Simulations are carried out from 1979 to 2012 over the Euro-Mediterranean area at 0.5° longitude/latitude resolution. We use the ISBA-A-gs LSM (Noilhan and Planton, 1989; Calvet et al., 1998) coupled with the CTRIP (standing for CNRM TRIP) river routing model (Decharme et al., 2010), an upgraded version of the TRIP model (Oki and Sud, 1997), to simulate the energy and water balances of the land surface as well as plant phenological development. ISBA-A-gs is a CO₂-responsive LSM that simulates vegetation dynamics and provides prognostic estimates of LAI (in the configuration described by Gibelin et al., 2006). It also features a multi-layer diffusion scheme for the soil water and energy balances (Decharme et al., 2013). The latter allows simulating soil moisture over a surface layer whose thickness is compatible with the generally acknowledged penetration depth of L-band microwave radiometers used to remotely sense SSM (10^{-2} – 10^{-1} m, according to Escorihuela et al., 2010, and Kerr et al., 2010).

To account for forcing uncertainties, LSM simulations are driven by four atmospheric datasets. Three of them are obtained from the ERA-Interim reanalysis (Dee et al., 2011) produced by European Centre for Medium Range Weather Forecasts (ECMWF; Reading, UK). ERA-Interim is the first dataset, while its two derivatives are: a version corrected using the Global Precipitation Climatology Project (GPCP) monthly precipitation data (Huffman et al., 2009); and WATCH Forcing Data ERA-Interim (WFDEI; Weedon et al., 2014) obtained by correcting precipitation, air temperature and downward short-wave (SW) radiation with monthly data from Climatic Research Unit (CRU; University of East Anglia, Norwich, UK; Harris et al., 2014). These three forcing datasets are characterised by the same large-scale synoptic variability and precipitation occurrences. The fourth dataset is chosen to have a LSM simulation independent from ERA-Interim: PGF (Sheffield et al., 2006) developed by the Terrestrial Hydrology Research Group at Princeton University (Princeton, NJ, USA).

The evaluation provides an overview of how atmospheric forcing uncertainties influence the simulation of surface hydrology. To frame uncertainties and their impacts, model input and output variables are compared to state-of-the-art reference datasets. Forcing data are compared to: E-OBS (van der Schrier et al., 2013) air temperature and precipitation; Global Precipitation Climatology Centre (GPCC; Deutscher Wetterdienst, Offenbach, Germany; Schneider et al., 2014) precipitation; CRU air temperature; NASA/GEWEX Surface Radiation Budget (SRB; Zhang et al., 2015) and Clouds and the Earth's Radiant Energy System (CERES; Wielicki et al., 1996) for long-wave (LW) and SW downward radiation. Some of these datasets are used to correct biases of atmospheric forcing (see Section 3.1) and thus are not independent references. Simulated LAI and SSM are compared to the remotely sensed Global Inventory Modeling and Mapping Studies (GIMMS; Zhu et al., 2013) and European Space Agency's Water Cycle Multi-mission Observation Strategy and Climate Change Initiative (ESA-CCI; Dorigo et al., 2014) datasets. River discharge simulations are evaluated using the recorded time series from 99 gauges of the Global Runoff Data Centre (GRDC, 2016) dataset.

The presented assessment involves the atmospheric conditions driving the simulation (forcing variables), the distributed land surface dynamics controlling fluxes at the interface with the atmosphere (LAI and SSM), and the integral of all catchment hydrological processes (river discharge). Moreover, we use long time series of remote sensing retrievals: GIMMS LAI data are available from 1981 to 2011; and ESA-CCI SSM covers the whole simulation period. Several studies evaluated the impacts of forcing uncertainties on large-scale LSM simulations: Guo et al. (2006) tested the sensitivity of soil moisture simulations forced by several meteorological forcing datasets; Decharme and Douville (2006b) and Szczypta et al. (2012) looked at the impact of forcing precipitation errors on river discharge simulations; Matera et al. (2010) analysed the sensitivity of simulated river discharge to changes in the atmospheric forcing data; Nasonova et al. (2011) investigated the impacts of forcing and surface parameter uncertainties on simulated runoff and evapotranspiration; Liu et al. (2016) ran multi-forcing and multi-model experiments to compare simulated and observed evapotranspiration. To our best knowledge, the extensiveness of the presented assessment exceeds those of previous studies focused on assessing LSM sensitivity to atmospheric forcing uncertainty. The large spectrum of comparisons and the length of the benchmark time series contribute to the completeness of the exercise.

The remainder of the paper is organised as follows. Sect. 2 describes the ISBA-A-gs LSM and CTRIP river routing model. Sect. 3 presents the atmospheric forcing datasets as well as all data sources used for evaluating the simulations. Sect. 4 outlines the experimental design of simulations and evaluation of model input and output. In Sect. 5 we show the results, which are discussed further in Sect. 6. Finally, Sect. 7 summarises the main findings of this study. The Supplement presents complementary results.

30 **2 SURFEX-CTRIP model**

To simulate the continental water balance and the related biophysical processes, we use the SURFEX (SURFace EXternalisée) version 7.3 modelling platform (Le Moigne, 2009) coupling the ISBA-A-gs LSM (Calvet et al., 1998) with

CTRIP through the OASIS3-MCT coupler (<https://verc.enes.org/oasis>). Given the atmospheric forcing, ISBA-A-gs computes the energy, water and carbon balances of spatially distributed independent columns of soil and vegetation. The resulting surface runoff and soil drainage are then routed by CTRIP through the river networks to simulate streamflow.

ISBA-A-gs is a CO₂-responsive version of the Interaction between Soil Biosphere Atmosphere (ISBA; Noilhan and Planton, 1989; Noilhan and Mahfouf, 1996) LSM simulating photosynthesis. It is based on the biochemical model of Goudriaan et al. (1985) modified by Jacobs et al. (1996) that describes the relation between photosynthesis and leaf stomatal aperture in the absence of water scarcity. ISBA-A-gs simulates two types of response to soil moisture stress for both herbaceous and forest vegetation (Calvet, 2000; Calvet et al., 2004): the drought-avoiding response increases water use efficiency at the emergence of the stress; while the drought-tolerant response decreases or keeps the water use efficiency stable throughout the stress period. ISBA parameters depend on soil and vegetation types. Land surface classifications are provided by the ECOCLIMAP-II/Europe database (Faroux et al., 2013) that covers Europe at 1 km resolution and is based on satellite observations.

We use the following ISBA-A-gs settings:

1. Coupling between above-ground biomass dynamics and soil water balance (“NIT”) (Calvet and Soussana, 2001; Gibelin et al., 2006).
2. Multilayer soil diffusion scheme explicitly solving the soil water and heat balance equations (“DF”) (Boone et al., 2000; Decharme et al., 2011; Decharme et al., 2013).

The basic hydrological unit of ISBA is a snow-vegetation-soil composite column. Each model grid cell consists of several homogeneous columns, to account for land surface heterogeneities at a finer spatial scale than the one imposed by the atmospheric forcing. **In this study we use 12 columns within 0.5° latitude/longitude grid cells. The size of model grid cells is determined by the atmospheric forcing spatial resolution.** The simulated water balance accounts for: canopy interception of precipitation; throughfall; snow accumulation and melting; infiltration; evapotranspiration; surface runoff; topsoil lateral subsurface flow; and soil drainage. The snowpack is simulated using a three-layer scheme (Boone and Etchevers, 2001). Soil ice is modelled according to Fuchs et al. (1978). Surface runoff fluxes due to both infiltration- and saturation-excess as well as topsoil lateral subsurface flow are modelled using subgrid approaches (Decharme and Douville, 2006a).

Heat and water balances are computed for each column at half-hourly time steps. The fluxes needed for coupling ISBA with other models are then aggregated by grid cell. Thus, daily grid-cell averages of surface runoff and deep soil drainage are used to force the CTRIP river routing model.

CTRIP routes the runoff and drainage flows from ISBA using two reservoirs per model grid cell (Decharme et al., 2010). Deep soil drainage feeds a linear groundwater reservoir, which in turn feeds the surface reservoir with an output flow proportional to its storage through a time constant.

The groundwater reservoir is not meant to represent actual groundwater processes, but to mimic the delay of soil drainage contributions to river discharge. As in Decharme et al. (2012), the groundwater reservoir time constant is fixed to 30 days.

The surface reservoir represents the segment of river channel contained in the grid cell. It is fed by groundwater flow and ISBA surface runoff from its grid cell, as well as by the outflows from the immediately upstream surface reservoirs. Its outflow is proportional to the ratio between channel storage and length through a variable flow velocity, which is given by Manning's formula (Arora and Boer, 1999). The CTRIP river channel parameters (length, width, roughness, and slope) are computed at 0.5° spatial resolution following Decharme et al. (2010), who also provide a detailed description of the model configuration used in this study.

3 Data

In this section we describe all datasets used as input and to evaluate model simulations, namely: atmospheric forcing and reference data (Sections 3.1 and 3.2); SSM and LAI remote sensing retrievals (Section 3.3); and river discharge measurements (Section 3.4).

The simulation grid covers the continental surfaces of the Euro-Mediterranean area (25°–75.5°N, 11.5°W–62.5°E) with a homogeneous 0.5° longitude/latitude resolution (Fig. 1).

3.1 Atmospheric forcing

ISBA-A-gs simulations were forced by 3-hourly time series of precipitation, downward LW and SW radiation, air temperature and humidity, wind speed, atmospheric pressure, and CO₂ concentration. The atmospheric datasets used to force model simulations are briefly described in the following list:

- ERA-I: ERA-Interim reanalysis (Dee et al., 2011) .
- P-ERA: ERA-I (Balsamo et al., 2015) whose precipitation amounts are bias-corrected so that their monthly values match those of GPCP data.
- WFDEI: WATCH Forcing Data ERA-Interim (Weedon et al., 2014). It is obtained from ERA-I by correcting precipitation, air temperature and downward SW radiation with monthly CRU data. Air temperature, pressure and humidity, as well as downward LW radiation are sequentially elevation-corrected. Moreover, air temperature is also corrected using mean monthly diurnal temperature ranges. Downward SW radiation is corrected for cloud cover and aerosol loading. Precipitation amounts are corrected to match CRU monthly averages.
- PGF: Global Meteorological Forcing Dataset for land surface modeling. It is based on the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR; Boulder, CO, USA; Kalnay et al., 1996) reanalysis. Precipitation and air temperature are bias-corrected so that monthly averages match those of CRU data; as for WFDEI, air temperature, pressure and humidity, as well as downward LW radiation are sequentially elevation-corrected; downward LW radiation is corrected with the NASA/GEWEX Surface Radiation Budget (SRB) monthly data (Zhang et al., 2013; Zhang et al., 2015); and downward SW radiation is corrected using both SRB and CRU data.

P-ERA and WFDEI inherit large-scale synoptic atmospheric circulation patterns and precipitation occurrences from ERA-I. Therefore, in the following we refer to ERA-I, ERA-P and WFDEI as ERA-based forcing. While WFDEI and PGF were available at 0.5° (latitude/longitude) horizontal resolution, ERA-I and P-ERA were interpolated from their original 0.75° resolution to 0.5° . All datasets span the experiment period 1979–2012. We used an experimental version of PGF, as the stable version available when simulations were run covers the period 1948–2008 (PGF, 2016). While ERA-based forcing provides rainfall and snowfall, PGF precipitation values were partitioned using an air temperature threshold of 1°C . The choice of the forcing datasets was motivated by two rationales. First, we wanted to compare the performance of a reanalysis product (ERA-I) with several bias-corrected versions. This type of atmospheric forcing is likely to be affected by biases, which in turn constitute a major source of uncertainty for land surface and hydrological simulations (Berg et al., 2003; Sheffield et al., 2004; Ngo-Duc et al., 2005; Weedon et al., 2011). Second, we included a forcing dataset based on a different reanalysis than ERA-I, i.e. PGF.

3.2 Atmospheric reference datasets

Reference datasets were processed for the following atmospheric forcing variables: precipitation, air temperature, and downward LW and SW radiation. To account for uncertainties in state-of-the-art climatological data, we used two datasets per variable.

Precipitation data were obtained from two station-based gridded datasets: the E-OBS version 11.0 product (Haylock et al., 2008; van der Schrier et al., 2013) covering the 1950–2015 period; and merging the 1900–2010 version 6 reanalysis and the 2011–2014 version 4 monitoring products of GPCC.

The air temperature datasets are: E-OBS, and CRU version 3.21 (see Sect. 3.1) spanning the 1901–2012 period.

The chosen reference datasets for downward LW and SW radiation monthly data are: SRB (see Sect. 3.1) versions 3.0 and 3.1 for SW and LW, respectively, covering the 1983–2007 period; and Clouds and the Earth's Radiant Energy System (CERES) version 2.8 from 2000 to 2014 (Wielicki et al., 1996).

E-OBS, GPCC reanalysis, and CRU datasets were available at 0.5° resolution. GPCC monitoring, SRB and CERES products were interpolated from the original 1° to 0.5° resolution via nearest-neighbour estimation.

3.3 Remote sensing retrievals

To evaluate simulated SSM and LAI, we used the ESA-CCI 1978–2014 dataset version 02, and the GIMMS 1981–2011 dataset version 1.3. These data sources were chosen since their temporal coverages are comparable to those of the forcing datasets. While the annual missing observation rates of GIMMS LAI are stationary, those of ESA-CCI SSM exhibit a decreasing trend due to the progressive inclusion of different microwave sensors. Despite its non-homogeneous time coverage, we decided to use ESA-CCI SSM for the whole simulation period (1979–2012) to perform a long term evaluation of LSM output.

3.3.1 ESA-CCI surface soil moisture

The ESA-CCI SSM product merges retrievals from several passive and active microwave satellite sensors (Liu et al., 2011; Dorigo et al., 2014). Although the sensing depth depends on soil moisture itself among other factors, microwave-based retrievals are regarded as informative of the moisture in the top few centimetres of soil (Escorihuela et al., 2010).

5 The source dataset consists of daily estimates (including missing values) spanning the 1978–2014 period at 0.25° horizontal resolution. First, daily maps were interpolated to the 0.5° model grid by averaging only over model pixels for which at least 50% of the possible retrievals were available. Retrievals flagged for insufficient quality were regarded as missing. As topographic relief is known to negatively affect remote sensing estimates of soil moisture (Mätzler and Standley, 2000), we discarded the time series for pixels whose average altitude exceeded 1500 m above sea level. Data on pixels with urban land
10 cover fractions larger than 15% were also discarded, to limit the effects of artificial surfaces. The altitude and urban area thresholds were set according to Draper et al. (2011) and Barbu et al. (2014), who processed ASCAT SSM retrievals for data assimilation exercises with the ISBA LSM.

The average time interval between two processed SSM values at a model pixel is 3.7 days. Splitting the domain at 50° N, retrievals are more frequent in the southern (on average every 3.1 days) than in the northern part (4.5 days). the lower
15 availability at high latitudes is partly explained by winter occurrences of frost and snow cover (Dorigo et al., 2014).

3.3.2 GIMMS leaf area index

The GIMMS LAI product is based on an artificial neural network algorithm. The algorithm was trained to map GIMMS normalized difference vegetation index (NDVI) to fraction of absorbed photosynthetically active radiation (FAPAR) and LAI retrieved by the Terra Moderate Resolution Imaging Spectroradiometer (MODIS; Yang et al., 2006). The training was
20 carried out on the overlapping 2000–2009 period. The algorithm was then used to generate LAI and FAPAR fortnightly estimates for the 1981–2011 period at 1/12° spatial resolution (Zhu et al., 2013).

As for ESA-CCI SSM, each retrieval map was interpolated only on those model pixels where missing data were below 50%. The average missing rate in processed data ranges from 34% in summer to 41% in winter, when snow cover hinders the retrievals in mountainous areas and in the northernmost part of the domain.

25 3.4 GRDC river discharge

To evaluate river discharge simulations, we used monthly gauge-based estimates provided by the Global Runoff Data Centre (GRDC, 2016). Ninety-nine GRDC gauges were paired with model pixels. For each gauge, 9 potential matching pixels were considered, namely the pixel containing the gauge and its 8 neighbours. The paired pixel was chosen as the one having the closest upstream drainage area to that of the gauge upstream catchment, provided that it fulfilled the following criteria:
30 minimum record length of 120 months; minimum upstream catchment area of $2 \cdot 10^4$ km²; maximum absolute difference of 10% between the catchment areas reported by GRDC and those described by the CTRIP river network. These criteria aim at

ensuring a meaningful comparison between observed and simulated values. Their enforcement is necessary for coping with the significant distortions in the model representation of the river network that are caused by the coarse spatial resolution.

Model pixels paired with GRDC gauges and the corresponding drained CTRIP river network are illustrated in Fig. 1. Catchment, location, upstream area and observation period of the 35 downstream gauges are reported in Fig. 2, whose rows

5 are indexed with the station numbers mapped in Fig. 1.

4 Experimental design

As our objective is the impact assessment of atmospheric forcing uncertainties on land surface hydrology, we performed a model simulation for each of the four considered forcing datasets (Sect. 3.1). All simulations span the 1979–2012 period, which is the longest overlapping time interval across the forcing datasets.

10 Model input and output are compared with a number of datasets, which are described in Sect. 3.2 to 3.4. Input values of precipitation, air temperature, as well as downward LW and SW radiation are compared to station- and satellite-based estimates. Among the model outputs, particular attention is dedicated to river discharge, SSM and LAI: river discharge can be seen as the integral of all catchment hydrological processes; SSM and LAI represent key components of the land surface biophysical state, as they are important controls on water, heat and carbon fluxes at the land-atmosphere interface. While
15 modelled SSM and LAI are compared to remote sensing retrievals, simulated river discharge time series are evaluated with gauge measurements. To have an overview of how the model simulates large scale hydrological processes, average annual cycles of forcing and simulated variables relevant to the water cycle are computed over large catchments.

ISBA-A-gs and CTRIP can produce outputs at half-hourly and daily time intervals, respectively. However, since three of the four forcing datasets are characterised by the same (ERA-I) synoptic variability, we choose to perform our analysis at the
20 monthly time scale. Indeed, we believe that a hydrological evaluation of the uncertainty deriving from the considered forcing datasets at (sub-)daily temporal scale may not add significant further information, because the ERA-based forcings have the same precipitation occurrence processes. Moreover, a daily time scale analysis may be severely affected by the lack of model representation of anthropic river regulation, which affects most major European rivers.

To compare model input/output variables with corresponding datasets, monthly series are computed using only the time
25 steps where reference data are available. Monthly anomalies are computed from monthly averages: at each grid-cell, the climatological mean and standard deviation are estimated for each calendar month. Then monthly averages are subtracted the climatological mean and divided by the climatological standard deviation to obtain normalised monthly anomalies. Anomaly statistics measure how well the model simulates departures from the annual cycle, i.e. the inter-annual variability.

4.1 Surface soil moisture

30 The water content of the ISBA top soil layer is compared to the ESA-CCI SSM remote sensing retrievals. We use the top layer, which is 1 cm thick, to take full advantage of the fine vertical soil discretisation of the DF scheme. The soil column is

represented with 14 layers over a 12 m depth following Decharme et al. (2013), who simulated soil moisture only over the root zone. Our choice is also based on the conclusions reached by previous studies. Escorihuela et al. (2010) studied the penetration depth of L-band microwave radiometry to estimate SSM: they found that brightness temperature, which is measured directly by the radiometer, was best correlated with soil moisture in the top 2 cm for dry soils and in the top 1 cm for wet soils. Dorigo et al. (2014), who validated ESA-CCI SSM data using ground-based measurements, used the records closest to the surface among those available within the top 10 cm.

To ensure comparability between model output and remotely sensed data, each SSM time series is linearly rescaled to the interval [0, 1] with respect to its minimum and maximum values. The resulting rescaled surface soil moisture (RSSM) can be regarded as a proxy of the topsoil saturation degree (Wagner et al., 1999; Albergel et al., 2012; Parrens et al., 2012; Polcher et al., 2016). This pixel-wise linear rescaling is designed to filter the effects of the discrepancies in the soil properties used as input by the LSM and the remote sensing retrieval algorithm. For instance, different soil composition maps may lead to different estimates of properties such as wilting point or field capacity and, in turn, to differences in soil moisture variability.

4.2 Leaf area index

Simulated LAI is compared to GIMMS data, which are available fortnightly. As the exact GIMMS estimate dates are unknown, monthly averages are computed from model values available on the 8th and 23rd days of each month. This choice is somewhat arbitrary and implies a 1 week uncertainty in the model dates for comparison. Zhu et al. (2013) estimated the root mean squared error of GIMMS LAI to be of the order of 0.5–0.9 m²m⁻². This uncertainty estimate is comparable to the average LAI monthly growth rates of both ISBA-A-gs and GIMMS over the study area (see the averaged LAI annual cycles in Fig. S2 to S5 of the Supplement). Thus, we believe our assumption would cause errors that are not larger than the uncertainties stemming from the GIMMS algorithm.

4.3 River discharge scores

Monthly river discharge simulations are compared to GRDC gauge measurements using several statistical scores. Relative errors of simulated mean (RE_{μ}) and standard deviation (RE_{σ}) account for biases and variability errors:

$$RE_{\theta} = \frac{\theta_{sim}}{\theta_{obs}} - 1 \quad (1)$$

where θ may be either μ or σ (standing for sample mean or standard deviation), while the subscripts *sim* and *obs* indicate whether the source data is simulated or observed. Correlation coefficients based on monthly time series (MC) quantify how accurately the simulations reproduce the shape and timing of measurements.

RE_{μ} , RE_{σ} and MC constitute a set of mathematically independent score functions quantifying the goodness of fit of simulated discharge time series. At their ideal values (RE_{μ} , $RE_{\sigma} = 0$ and $MC = 1$), the simulated time series is identical to the measured one. Moreover they account for errors in modelling the first three fundamental hydrological functions of a watershed as

identified by Black (1997), Wagener et al. (2007) and Yilmaz et al. (2008): the overall water balance resulting from the partition of precipitation between evapotranspiration and infiltration plus direct runoff (RE_{μ}); the soil water storage dynamics distributing the excess precipitation among faster and slower runoff components (RE_{σ}); and the water release and routing processes determining timing and shape of the hydrograph (MC).

- 5 To summarise the information of RE_{μ} , RE_{σ} and MC with a single aggregate score, we compute the Kling-Gupta efficiency (KGE; Gupta et al., 2009):

$$KGE = 1 - \sqrt{RE_{\mu}^2 + RE_{\sigma}^2 + (1 - MC)^2} \quad (2)$$

which is the Euclidean distance from the ideal point in the [RE_{μ} , RE_{σ} , MC] score space. We prefer KGE to other widely used residual-based summary statistics such as the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) or the related mean squared error (MSE), to prevent the following drawbacks identified by Gupta et al. (2009): in catchments with high discharge variability, NSE and MSE overrate model simulations with large biases; and if MC is smaller than 1 (always in real cases), simulations underestimating discharge variability are overrated. While NSE = 0 means that the model is a worse predictor than the observed average, negative KGE values have no defined interpretation. This advantage of NSE, however, is significant in the domain of poor model performance.

- 15 An aggregate score is often used as unique criterion in model calibration, because it allows applying efficient single-objective optimisation algorithms. However, model calibration is an inherently multi-objective problem (Gupta et al., 1998). Thus, aggregate scores are not as informative as their individual components for evaluating model performance (Gupta et al., 2009).

To evaluate the model ability to capture the inter-annual variability at the monthly scale, we compute correlations between anomaly time series (AC). AC are based on normalised departures from the averaged discharge annual cycle. Thus they allow filtering out the correlation induced by the climatological seasonality, which contributes significantly to MC.

4.4 Catchment averaged annual cycles

We compute catchment averaged annual cycles for the following forcing variables: downward LW and SW radiation, air temperature at 2 m, wind speed, relative air humidity, and total precipitation. Of the simulated land surface states and fluxes, we show: river discharge, evapotranspiration, LAI, and RSSM. To be consistent with river discharge, which is the integral of upstream hydrological processes, annual cycles are computed using only the time steps when river discharge measurements are available.

5 Results

The evaluation of model simulations against the datasets described in Sect. 3 is reported in the following. Catchment averaged simulated annual cycles of hydrologically relevant variables are described in the Supplement for four river basins (Danube, Rhone, Ebro and Po).

5.1 Surface soil moisture

We compare the simulated water content of the top soil layer of the ISBA model with the ESA-CCI SSM remote sensing retrievals (Sect. 3.3.1). To ensure comparability, each SSM time series is linearly rescaled with respect to its minimum and maximum, as motivated in Sect. 4.1.

Global correlation values are computed using the entire datasets and are reported in the upper-left corners of the maps in Fig. 3. As expected, similar scores are obtained for the different simulations, among which the PGF-forced SSM is slightly less correlated with ESA-CCI data. Looking at this small difference, we must recall that the atmospheric synoptic variability of PGF is different from that of ERA-based forcing. ERA-I, P-ERA and WFDEI are characterised by the same precipitation occurrences, while the intensities of individual precipitation events may differ according to the applied bias-correction (Sect. 3.1).

Temporal correlation maps between simulated and remotely sensed SSM are computed for both monthly absolute and anomaly time series (Fig. 3). Pixel-wise temporal correlation is not affected by the rescaling. Similar correlation patterns emerge across the simulations. Monthly time series are most correlated (between 0.5 and 1) in southern and western Europe, northern Africa, Middle East, and the portion of central Asia included in the domain. Smaller positive correlations (below 0.5) characterise large areas of central and eastern Europe, the Caucasus, and south-eastern Russia. Null or negative correlations are found in Scandinavia, eastern Baltic and northern Russian regions, as well in mountainous areas such as the Alps and the Carpathians. Possible causes for these low values may be: vegetation density, topography and soil frost occurrences, all of which are known to negatively affect SSM retrievals; and model misrepresentation of surface hydrological processes that are particularly relevant in cold and mountainous regions. Despite generally smaller values, temporal correlations of anomalies exhibit map patterns that are similar to those described for monthly values.

Fig. 3 takes the ERA-I correlation as a reference to show the departures of the other forcing datasets. Consistently with the above-mentioned similarities between correlation maps, absolute differences are small: they are below 0.1 for at least 90% (monthly averages) and 74% (anomalies) of the mapped grid-cells. Over 94% of P-ERA monthly and anomaly correlations are within a 0.05 range of ERA-I values, with small positive differences scattered in the eastern part of the domain. The same is true for 81% (70%) and 66% (44%) of, respectively, WFDEI and PGF monthly values (anomalies). WFDEI negative departures are mostly located in northern Africa. PGF yields the largest differences, particularly those of negative sign.

Correlation coefficients computed separately for each calendar month (Fig. 4) report the seasonal variations in the agreement between simulated and remotely sensed RSSM. The correlation annual cycles of monthly absolute values reach their peaks

from June to September and their minima in November-December and March-April, indicating that RSSM maps are more consistent during the driest months. All monthly anomaly correlations have their maxima in September and their minima from December to March. Thus the agreement between deviations from RSSM climatologies tend to steadily increase through summer, peaking when soils are driest. While PGF correlations are systematically the smallest for both monthly absolute and anomaly values, the differences are not dramatic and their cycle shapes are similar to those of ERA-based data. The similarities of (R)SSM correlation statistics across the simulations suggest that the model reacts robustly to forcing uncertainty. However, it may also indicate that the discrepancies found between modelled and remotely sensed (R)SSM could be caused by the LSM representations of physical processes and the remote sensing retrieval algorithms. The correlation annual cycles (Fig. 4) point in the same direction. Indeed, correlations are largest in summer when SSM variability is minimum due to dry conditions. Instead they are smallest in winter, when SSM variability is maximum due to precipitation and thus the description of SSM dynamics becomes more important.

5.2 Leaf area index

Simulated phenology is compared to the satellite based GIMMS LAI product (Sect. 3.3.2).

Temporal correlations of monthly averages are large (above 0.7) over more than 75% of the pixels where data are available, for all forcing datasets (Fig. 5). The largest values are mostly in the central and eastern parts of the domain. Very small or even negative correlations are scattered across the southern part. Irrigation, which is not simulated by ISBA-CTRIP, may be partially accountable for these discrepancies in some areas, as for instance the Nile delta and Mesopotamia.

Anomalies are systematically less temporally correlated than monthly values in all simulations. Over 90% of pixels with data availability score values below 0.5. However, values below 0.1 are less than 3%, so the areas with null or negative correlations are very limited. Consistently with temporal correlation maps, global correlation coefficients (upper left corners of maps in Fig. 5) are large for monthly averages (0.79–0.81) and significantly smaller for anomalies (0.32–0.34).

Fig. 5 maps ERA-I temporal correlation coefficients and the differences yielded by the other simulations, as done for SSM (Fig. 3). Discrepancies between correlation maps of different simulations are generally small. Over 94% and 89% of the absolute differences with ERA-I correlations are smaller than 0.1 for monthly values and anomalies, respectively. P-ERA yields the most similar maps to ERA-I: more than 95% and 83% of monthly and anomaly correlation differences are below 0.05. This is expected, as P-ERA is identical to ERA-I except for precipitation amounts. Consistently, WFDEI and PGF, which feature progressively increasing divergences from ERA-I forcing, yield slightly larger correlation differences. These are below 0.05 for 82% (69%) and 81% (61%) for monthly (anomaly) correlations of WFDEI and PGF. While differences from ERA-I correlations are prevalently positive for P-ERA, their signs are more mixed for the other two forcing datasets. The largest clearly distinguishable difference signals, which are yielded by WFDEI and PGF monthly correlations, are located along the Caucasus and Pontic mountain ranges.

LAI correlation statistics show a general agreement across forcing datasets. This is also true for monthly correlation coefficients (Fig. 6), which follow very distinguishable seasonal patterns. Correlations between monthly averages peak in

July-August (0.85), when LAI reaches its maximum. Instead, anomaly correlation are largest (0.50–0.55) in April–May, when the phenological development rate is likely to be maximum. Both correlation annual cycles are lowest in winter, when the ISBA-A-gs dynamic vegetation model sets LAI of several vegetation types to prescribed values until the next leaf onset occurs triggering a new growing season. This does not apply to evergreen trees and winter crops. The prescribed minimum
5 LAI values may be partly accountable for the small winter correlations. They may have a particularly negative impact on anomalies by hindering any winter inter-annual variability.

All correlation statistics are very stable across the forcing datasets. This is similar to what observed for SSM, with two exceptions: PGF statistics are more aligned with the other forcing datasets; and correlations for monthly averages are generally larger. Thus, compared to SSM, LAI correlation statistics are less sensitive to forcing uncertainty. This is not
10 surprising as the impact of precipitation uncertainty is larger on SSM than on LAI. Indeed, LAI dynamics are strongly influenced by the periodicity of plant development and by forcing variables that are less uncertain than precipitation, such as air temperature and SW radiation. In other words, the LAI degrees of freedom are reduced compared to SSM. Furthermore, LAI is less sensitive than SSM to short-term uncertainties in the atmospheric forcing, due to its larger characteristic time scale.

15 **5.3 River discharge scores**

Simulated monthly river discharge time series are evaluated using the GRDC dataset (Sect. 3.4). We use the statistical scores defined in Sect. 4.3: RE_{μ} , RE_{σ} , MC, AC, and KGE. Score values are summarized in exceedance frequency plots (Fig. 7) for all 99 selected GRDC gauges. Moreover, they are reported in detail for the 35 most downstream gauges (Tables 1 and 2).

To facilitate graphical display and interpretation, Fig. 7 plots the absolute values of RE_{μ} and RE_{σ} . The abscissae of $|RE_{\mu}|$ and
20 $|RE_{\sigma}|$ are inverted, so that all exceedance frequency plots can be read in the same way: the higher the curve, the better the performance.

The simulations yield similar MC and AC summary curves (Fig. 7). Approximately 60% of the gauges have MC and AC values above 0.6. The WFDEI and P-ERA simulations slightly outperform the others in terms of MC and AC, respectively, but differences are small. The inter-forcing spread becomes larger if we look at relative errors. In terms of RE_{μ} , ERA-I
25 performs best followed by PGF, WFDEI and P-ERA. While $|RE_{\mu}|$ is below 0.2 for over 70% of ERA-I-forced time series, the same happens for less than 50% of P-ERA series. The spread is even wider for $|RE_{\sigma}|$. The PGF simulation dominates almost the entire frequency spectrum, followed by WFDEI, ERA-I and P-ERA. More than 60% of PGF and less than 25% of P-ERA series have $|RE_{\sigma}|$ below 0.4. The $|RE_{\mu}|$ and $|RE_{\sigma}|$ exceedance frequency curves of P-ERA are almost systematically dominated by those of the other simulations. The relative errors (in particular RE_{σ}) have a strong impact on the summary
30 curves of the aggregate score KGE, according to which the P-ERA simulation is dominated at almost all frequencies. Instead, a clear KGE ranking cannot be established among the other three simulations. **While KGE does not provide diagnostic information on the causes of better or worse model performance, the individual scores show that forcing**

uncertainties have larger impacts on the mean and standard deviation of the simulated discharge than on shape, timing and inter-annual variability.

The spatial patterns of river discharge scores are mapped and discussed in the Supplement (Fig. S1). The main findings are summarised here: the similarities between correlation maps indicate that the timing and shape of simulated monthly discharge are relatively insensitive to forcing uncertainty; relative error maps are also similar, but larger discrepancies emerge across simulations; in particular, P-ERA scores the largest $|RE_\sigma|$ over several catchments (consistently with Fig. 7); KGE spatial patterns are largely influenced by those of relative errors, in particular $|RE_\sigma|$.

River discharge scores computed at gauges having downstream/upstream connections may carry redundant information. Thus simulations performing systematically better/worse than others in multi-gauged catchments may be overrated/underrated, as such catchments would be virtually assigned larger weights when pooling the scores for comparison (Fig. 7 and S1). To avoid redundancy, we look at river discharge scores at the most downstream gauges of each catchment (Tables 1 and 2). Discharge series at these gauges are the closest estimates of surface water flowing into the oceans. Thus, detailed assessments of discharge simulations at these river cross-sections are of crucial importance, both to validate the simulated water balance of large catchments and for the integration with coastal water and ocean models.

All simulations overestimate the standard deviation ($RE_\sigma > 0$) of discharge time series at the majority of downstream gauges (Table 1). Systematic underestimation occurs only for Severnaya Dvina, Mezen and Torneaelven rivers, which drain sub-Arctic catchments. At 30 of 35 downstream gauges the P-ERA simulation overestimates average discharge ($RE_\mu > 0$), while for other simulations negative and positive biases are fairly balanced. In terms of RE_μ and RE_σ , P-ERA is inferior to other simulations at 20 and 27 downstream gauges, respectively.

MC is above 0.5 at over half of the downstream gauges for all simulations (Table 1). The same applies to AC, whose values are not systematically inferior to MC (Table 2): for simulated time series with MC above 0.5, AC tends to be slightly smaller than MC; the opposite happens for MC below 0.5. Catchments with larger MC and AC values are located in central and western Europe (Danube, Rhone, Tejo, Weser, Meuse), as also shown in Fig. S1. Relatively low scores characterise some northern and eastern European catchments (Volga, Don, Neva, Ural, Vuoksi, Kovda, Luleaelven). P-ERA yields the best MC values for more than a third of the gauges. Both ERA-I and PGF, which perform better than the other simulations in terms of RE_μ and RE_σ , produce the worst MC values for over a third of the gauges. Similar counts are observed for AC as well.

The P-ERA simulation yields the worst KGE at 25 downstream gauges (Table 2). Of these gauges, for 20 it also delivers the worst RE_μ and RE_σ scores; for 8, it yields the best MC; and the worst MC only for 3. This is in agreement with the large influence of RE_μ and RE_σ on KGE already found when discussing Fig. 7 and S1. Moreover, this is caused by forcing uncertainty having the largest impacts on the mean and standard deviation of simulated discharge. Instead, shape, timing and inter-annual variability (through anomaly correlation) are less sensitive to forcing uncertainty.

5.4 Catchment averaged annual cycles

To have an overview of how atmospheric forcing uncertainty affects simulated hydrology at catchment scale, we compare the average annual cycles of forcing and LSM output variables over four catchments: Danube, Rhone, Ebro and Po. These catchments were chosen as they contribute to the water balance of the Mediterranean sea, even if indirectly as the Danube; and because the model reproduces fairly well discharge measurements at their downstream GRDC gauges (Tables 1 and 2). The computed catchment averaged annual cycles are shown and described in detail in Sect. S2 of the Supplement (Fig. S2 to S5). The main findings are summarised in the following paragraphs.

Precipitation features the most different annual cycles across atmospheric datasets. While forcing values are often larger than measurement-based averages, the seasonal distributions are generally in good agreement within each catchment.

Simulated annual cycles of RSSM are close. The pixel-wise rescaling (Sect. 4.1) is partly accountable for the similarities, which are particularly strong among ERA-based simulations. Simulated cycles are in phase with ESA-CCI RSSM, but their amplitudes are generally overestimated. This may be due to structural differences between modelled and remotely sensed SSM that are beyond first-order soil property discrepancies, which should have been filtered out by the rescaling.

Simulated LAI cycles are relatively insensitive to forcing uncertainty. In the Danube, Rhone and Po catchments, they are in good agreement with GIMMS LAI when phenological development is fastest, but they overestimate annual maxima and during senescence. In the Ebro catchment, simulated cycles overestimate the remote sensing-based one.

At most downstream gauges, shape and timing of observed discharge cycles are well reproduced. Simulated cycle amplitudes are generally larger than for measurements: while summer-autumn discharge is underestimated, winter and spring values are overestimated and the inter-simulation spread is maximum, consistently with the precipitation cycles.

20 6 Discussion

To evaluate how atmospheric forcing uncertainties affect the simulation of land surface hydrological processes, we compared model input and output with several independent datasets. In the following subsections, we discuss the main results and we draw some recommendations for future research.

6.1 Atmospheric data

25 Precipitation is the most uncertain forcing variable (Fig. S2 to S5). Conversely, the least uncertain forcing variable is air temperature, followed by SW and LW radiation. Both P-ERA and WFDEI forcing datasets are derived from ERA-I. While P-ERA differs from ERA-I only because of monthly precipitation amounts, WFDEI also features differences in SW radiation, air temperature and relative humidity. PGF is independent from ERA-based datasets. In addition to precipitation, also wind speed and relative air humidity show remarkable discrepancies between PGF and the other datasets.

30 **Limitations of (and alternatives to) bias correction methods adjusting a sub-set of the forcing variables or not involving statistical moments beyond the mean are discussed by Haddeland et al. (2012) and Sippel et al. (2016).**

Some of the used reference atmospheric datasets are not independent from the forcing. For example, CRU air temperature is used as reference and to bias-correct WFDEI and PGF forcing. The extensive use of CRU data is motivated by its relatively high update frequency and resolution (0.5°), which make this dataset attractive for land surface modelling. Moreover, since PGF and WFDEI are based on different reanalyses, their bias corrections yield different sub-diurnal variabilities. However, to reduce dataset interdependencies, future work on the evaluation of forcing uncertainties may benefit from using a wider ensemble of state of the art meteorological datasets, among others: the Global Soil Wetness Project (GSWP3) forcing dataset (Yoshimura and Kanamitsu, 2013); the Goddard Institute for Space Studies Temperature (GISSTEMP) analysis (Hansen et al., 2010); the Global Historical Climatology Network temperature dataset (Lawrimore et al., 2011); The Berkeley Earth Surface Temperatures (BEST) dataset (Rohde et al., 2013); and the Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset (Beck et al., 2017b).

6.2 Surface soil moisture

The similarities in correlation statistics across simulations suggest that modelled SSM is rather insensitive to forcing uncertainty (Fig. 3 and 4). It also suggests, however, that some causes of the discrepancies between modelled and remotely sensed SSM may be rooted in the representations of the physical processes underlying LSMs and remote sensing retrieval algorithms. Advancing the understanding of these discrepancies is crucial for assimilating remotely sensed SSM into LSMs (Reichle et al., 2008; Draper et al., 2012; Carrera et al., 2015; Barbu et al., 2014; Fairbairn et al., 2015). In light of the presented results, we suggest the following research directions: joint assessments of forcing and model uncertainty (Entin et al., 1999; Guo et al., 2006; Matera et al., 2010; Nasonova et al., 2011); extension of LSM-satellite comparisons to land surface state variables and fluxes driving model simulation or remote sensing retrieval of SSM, e.g. soil brightness temperature (Parrens et al., 2014; Muñoz-Sabater, 2015; Barella-Ortiz et al., 2017) or radar backscattering coefficient (Stoffelen et al., 2017); investigating the improvement or inclusion in LSMs of physical processes that are relevant to topsoil hydrology at high latitudes, e.g. snowmelt, flooding and ponding (Gouttevin et al., 2013).

The SSM statistics yielded by PGF show a slight departure from the range of ERA-based simulations. As PGF is the only forcing dataset not characterised by ERA-I synoptic variability, this result raises questions about the impact of precipitation occurrences on SSM correlations compared to other forcing variables. Indeed, SSM is likely to be very sensitive to precipitation occurrences, as the modelled/observed surface soil layer is very thin (less than 10 cm). Following investigation steps in this regard should be carried out at finer temporal scales than monthly, at which higher frequency variability is filtered out by the time averaging. For instance, daily correlations would be more informative about the consistency of forcing precipitation events. Thus one could test SSM sensitivity to precipitation occurrences compared to amounts. For this, LSMs should be forced by several atmospheric datasets with independent precipitation occurrences. As in this study 3 of 4 simulations are forced by the same precipitation occurrences, we analysed the output at the monthly time scale.

The ESA-CCI RSSM annual cycles display smaller amplitudes than the simulations, whose cycles are very close. These observations are valid not only for the catchments analysed in Sect. S2 (Fig. S2 to S5), but for all the areas drained by the

chosen 35 downstream GRDC gauges (not shown). We may draw a connection between this result and the spectral analysis performed by Polcher et al. (2016), who compared SSM simulated by the ORCHIDEE LSM (Krinner et al., 2005) with SMOS retrievals (Kerr et al., 2010) in the Iberian Peninsula. They found the power ratio of low (seasonal to annual) to high (daily to monthly) frequencies in the SSM signal to be significantly smaller for SMOS than for the LSM. To validate this hypothesis, further large-scale analyses of SSM signals may be needed, also involving several models and forcing datasets to account for as many uncertainty sources as possible.

The rescaling (Sect. 4.1) ensuring comparability between modelled and remotely sensed SSM may be an additional source of uncertainty. For future work, we recommend testing more robust rescaling techniques. For example, one could use physically meaningful parameters such as wilting point and field capacity instead of the time series extreme values, as done here.

6.3 Leaf area index

Compared to SSM, simulated monthly LAI is generally more correlated with the remotely sensed estimates; and LAI correlation statistics are more insensitive to forcing uncertainty (Fig. 5 and 6). These differences may be due to the larger impact of precipitation uncertainty on SSM than on LAI. Simulated LAI is relatively insensitive to atmospheric forcing uncertainties, suggesting that phenological dynamics are reproduced robustly by the model.

The correlation between simulated and remotely sensed LAI is largest during the maximum phenological development phase (Fig. 6). This is an encouraging result for the simulation of crop yield and, in general, of the primary production of land surface ecosystems. LAI correlations reach their minima in winter. As ISBA-A-gs prescribes a minimum LAI value for non-growing seasons, future research could evaluate the impact of this assumption on winter correlations and inter-annual variability by means of tailored sensitivity analyses.

Simulated catchment averaged annual cycles of LAI feature systematic overestimations at the maximum phenological development and during senescence (Fig. S2 to S5). This is consistent with the findings of Zhu et al. (2013), who compared the GIMMS LAI annual cycles with an ensemble of 18 Earth system models (see Sect. 4 and Fig. 7 of the cited article). They found GIMMS LAI to be generally below the models mean by approximately the ensemble standard deviation. Moreover, they pose the question of whether “dynamic vegetation models overestimate carbon fixation and/or allocation of biomass to leaves”.

Validation studies on the biomass production simulated by ISBA-A-gs have been performed over agricultural sites in France (Calvet et al., 2012; Canal et al., 2014; Dewaele et al., 2017). Further investigation may be carried out by extending the spatial scale of the validation (see e.g. Smith et al., 2010a,b). Similarly, irrigation modelling and its impacts on simulated agricultural production need to be validated at large spatial scales: Garrigues et al. (2015) evaluated how ISBA-A-gs simulated evapotranspiration over an irrigated Mediterranean agricultural site. Dynamic vegetation modelling and realistic representations of irrigation are crucial to predict crop yield and assess the sustainability of crop water management (Jägermeyr et al., 2016).

6.4 River discharge

Simulated monthly river discharge is generally well correlated with GRDC estimates. Anomaly correlations are comparable to those obtained for the monthly time series. Both MC and AC have similar distributions and spatial patterns across the simulations, indicating that the shape and timing of simulated discharge time series are relatively insensitive to forcing uncertainty (Fig. 7 and S1). The performance spread between simulations becomes larger when we consider RE_{μ} or RE_{σ} . The P-ERA forced simulation produces the largest relative errors at the majority of the gauges, while ERA-I and PGF yield the best overall results in terms of $|RE_{\mu}|$ and $|RE_{\sigma}|$ respectively (Fig. 7 and Table 1).

Positive and negative biases are well balanced for all simulations except for P-ERA that overestimates average discharge at most downstream gauges (Table 1). In contrast, all simulations tend to overestimate the standard deviation of the measured time series. This overestimation is in agreement with the generally larger amplitude of simulated discharge annual cycles compared to the measurements (Fig. S2 to S4). These results suggest that forcing uncertainty affects more the mean and standard deviation rather than the timing, shape and inter-annual variability of simulated monthly discharge.

The seasonal distributions of measured and simulated discharge are often different. In particular, simulations tend to underestimate summer-autumn discharge that correspond to the low flow conditions in the study area. Furthermore, in several catchments simulated winter-spring discharge peaks are significantly larger than the corresponding reference data. These systematic errors, which are found in all simulations, may be partly due the lack of model representations of physical processes and anthropic alterations affecting the streamflow regime.

In the past few years, new components have been tested to provide the ISBA-CTRIIP modelling suite with a progressively more complete representation of hydrological processes. Vergnes and Decharme (2012) implemented a simple groundwater scheme coupled to the river channel in each model grid cell allowing bidirectional water exchanges through the riverbed, showing improvements in the simulations of total water storage variations and river discharge. Vergnes et al. (2014) introduced the representation of capillary water rise from the groundwater reservoir, which provides the lower boundary condition for the moisture redistribution in the unsaturated soil column. To try to improve the modelled discharge response to rainfall events, the current Richards equations representation of water movement in unsaturated soil might be complemented with other parametrisations as, for instance, preferential flow in macro-pores (Beven and Germann, 2013).

Regulation of lakes and artificial reservoirs has a large impact on river discharge (Biemans et al., 2011; Zhou et al., 2016). Its representation in LSMs and hydrological models at global scale has proven to be beneficial to river discharge and irrigation modelling (Hanasaki et al., 2006; Pokhrel et al., 2012). Moreover it allows simulating the impacts of anthropic regulation on water-related activities. Further work should be carried out to embed reservoir regulation modules in LSMs.

River discharge is the integral of all upstream hydrological processes. Land water storage is a fundamental state variable for determining the catchment response to meteorological forcing. While large-scale direct observations are not available, the Gravity Recovery and Climate Experiment (GRACE) satellite mission provides land water storage variation estimates derived from measurements of the terrestrial gravity field variations (Swenson et al., 2003). Several studies have compared

LSM simulated land water storage to GRACE estimates (see e.g. Alkama et al., 2010; Becker et al., 2011; Grippa et al., 2011; Vergnes and Decharme, 2012). If integrated with the evaluation of other components of the land water cycle, these comparisons could help back-tracking the error sources of LSM simulations.

7 Conclusions

5 To assess the hydrological impacts of atmospheric forcing uncertainties, we compared land surface model (LSM) simulations forced by several meteorological datasets against station-based and remote sensing estimates. Simulations were run for the 1979–2012 period over the Euro-Mediterranean area, which is widely acknowledged for its drought vulnerability especially under climate change scenarios.

We used the ISBA-CTRIP modelling platform, which simulates dynamic plant above-ground biomass and a multi-layer
10 diffusion scheme for the unsaturated soil water and energy balances, together with river discharge, a non-prognostic linear groundwater reservoir, and a non-linear surface reservoir simulating variable-velocity river channel flow.

The simulations were forced using four atmospheric datasets and the evaluation of model outputs focused on surface soil moisture (SSM), leaf area index (LAI), and river discharge.

Simulated and remotely sensed annual cycles of SSM are generally in phase, but simulated amplitudes are systematically
15 larger despite the rescaling. The comparisons indicate that SSM simulations are relatively insensitive to forcing uncertainty. At the same time, the differences between simulated and remotely sensed SSM may be due to structural inconsistencies between the assumptions used by the LSM and the retrieval algorithm.

LAI temporal correlation maps as well as global correlation coefficients show a good agreement between modelled and
20 observed values for monthly averages. Systematically smaller values are obtained for monthly anomalies. Simulated LAI is most correlated with remotely sensed data during the typical seasons of maximum phenological development (spring and summer). The lower winter correlations may be negatively affected by the minimum LAI values prescribed by ISBA-A-gs during non-growing seasons. In general, LAI correlation statistics are at least as insensitive to forcing uncertainty as those of SSM.

Precipitation is the most uncertain forcing variable. Nevertheless, the timings of precipitation annual cycles are generally
25 similar within each catchment. The least uncertain forcing variables are air temperature, LW and SW radiation.

The annual cycles of simulated discharge have generally larger amplitudes compared to GRDC records. This translates into almost systematic overestimation of winter-spring high flows and underestimation of summer-autumn low flows. Positive and negative errors in long term average discharge are fairly balanced for all simulations except for P-ERA, which yields positive biases at most gauges. P-ERA also produces the largest standard deviation errors at the majority of the gauges, while
30 ERA-I and PGF yield the best overall results in terms of bias and standard deviation error, respectively. The spread between simulations is significantly reduced when considering correlations of monthly discharge and anomalies. Moreover, anomaly correlations are not inferior to those of raw monthly values, thus highlighting the model ability to reproduce the inter-annual

variability of river discharge. The impacts of forcing uncertainty are larger on the mean and standard deviation rather than the timing, shape and inter-annual variability of simulated discharge. Moreover, the errors in the seasonal distribution may be due to the lack of model representations of physical processes and anthropic alterations that affect streamflow.

Several results point at structural discrepancies between simulations and reference datasets that seem not to be addressable to forcing uncertainty. Therefore, we identify several research directions to improve LSMs and to foster the assimilation of remote sensing retrievals. The differences between simulated and remotely sensed SSM may be further investigated by extensively comparing simulated and retrieved land surface variables driving SSM dynamics, such as soil brightness temperature or radar backscattering coefficient (see e.g. Parrens et al., 2014; Stoffelen et al., 2017). To improve the simulation of freshwater availability, LSMs may progressively integrate more physically based groundwater schemes (see e.g. Vergnes and Decharme, 2012; Vergnes et al., 2014) as well as lake and reservoir regulation (see e.g. Hanasaki et al., 2006; Pokhrel et al., 2012). Moreover, simulated plant biomass growth should be validated at large spatial scales using agricultural statistics (see e.g. Smith et al., 2010a,b). In general, to better identify the sources of uncertainty in LSM simulations, forcing and model uncertainties may be assessed jointly via extensive multi-forcing and -model experiments (see e.g. Nasonova et al., 2011).

15 Acknowledgements

We are thankful to Emanuel Dutra and Gianpaolo Balsamo (European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK) for providing the ERA-Interim atmospheric forcing dataset as well as the GPCP-corrected version. The work of Emiliano Gelati was supported by the French REMEMBER project (ANR 2012 SOC&ENV 001) within the HYMEX initiative. The work of Marie Minvielle and of David Fairbairn was supported by the following European FP7 projects: earthH2Observe (grant agreement 603608) and ImagineS (grant agreement 311766), respectively. Graham P. Weedon was supported by the Joint DECC and Defra Integrated Climate Program - DECC/Defra (GA01101).

References

- Albergel, C., de Rosnay, P., Gruhier, C., Muñoz-Sabater, J., Hasenauer, S., Isaksen, L., Kerr, Y., Wagner, W.: Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations, *Remote Sens. Environ.*, 118, 15, 215–226, <https://doi.org/10.1016/j.rse.2011.11.017>, 2012.
- Albergel, C., Dorigo, W., Balsamo, G., Muñoz-Sabater, J., de Rosnay, P., Isaksen, L., Brocca, L., de Jeu, R., and Wagner, W.: Monitoring multi-decadal satellite earth observation of soil moisture products through land surface reanalyses, *Remote Sens. Environ.*, 138, 77–89, <https://doi.org/10.1016/j.rse.2013.07.009>, 2013a.

- Albergel, C., Dorigo, W., Reichle, R. H., Balsamo, G., de Rosnay, P., Muñoz-Sabater, J., Isaksen, L., de Jeu, R., and Wagner, W.: Skill and global trend analysis of soil moisture from reanalyses and microwave remote sensing, *J. Hydrometeorol.*, 14, 1259–1277, <https://doi.org/10.1175/JHM-D-12-0161.1>, 2013b.
- Albergel, C., Munier, S., Leroux, D. J., Dewaele, H., Fairbairn, D., Barbu, A. L., Gelati, E., Dorigo, W., Faroux, S., Meurey, C., Le Moigne, P., Decharme, B., Mahfouf, J.-F., Calvet, J.-C.: Sequential assimilation of satellite-derived vegetation and soil moisture products using SURFEX_v8.0: LDAS-Monde assessment over the Euro-Mediterranean area, *Geosci. Model Dev.*, 10, 3889–3912, <https://doi.org/10.5194/gmd-10-3889-2017>, 2017.
- Alkama, R., Decharme, B., Douville, H., Becker, M., Cazenave, A., Sheffield, J., Voldoire, A., Tyteca, S., and Le Moigne, P.: Global evaluation of the isba-trip continental hydrological system. Part I: comparison to GRACE terrestrial water storage estimates and in situ river discharges, *J. Hydrometeorol.*, 11, 583–600, <https://doi.org/10.1175/2010JHM1211.1>, 2010.
- Andréassian, V., Le Moine, N., Perrin, C., Ramos, M. H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: the case of calibrating hydrological models, *Hydrological Processes*, 26, 2206–2210, <https://doi.org/10.1002/hyp.9264>, 2012.
- Arora, V. K., and Boer, G. J.: A variable velocity flow routing algorithm for GCMs, *J. Geophys. Res.*, 104, 30965–30979, 1999.
- Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land surface reanalysis data set, *Hydrol. Earth Syst. Sci.*, 19, 389–407, <https://doi.org/10.5194/hess-19-389-2015>, 2015.
- Barbu, A. L., Calvet, J.-C., Mahfouf, J.-F., and Lafont, S.: Integrating ASCAT surface soil moisture and GEOV1 leaf area index into the SURFEX modelling platform: a land data assimilation application over France, *Hydrol. Earth Syst. Sci.*, 18, 173–192, <https://doi.org/10.5194/hess-18-173-2014>, 2014.
- Barella-Ortiz, A., Polcher, J., de Rosnay, P., Piles, M., and Gelati, E.: Comparison of measured brightness temperatures from SMOS with modelled ones from ORCHIDEE and H-TESEL over the Iberian Peninsula, *Hydrol. Earth Syst. Sci.*, 21, 357–375, <https://doi.org/10.5194/hess-21-357-2017>, 2017.
- Bastidas, L. A.: Parameter estimation for hydrometeorological models using multi-criteria methods, PhD dissertation, Department of Hydrology and Water Resources, University of Arizona, Tucson, 1998.
- Bastidas, L. A., Hogue, T. S., Sorooshian, S., Gupta, H. V., and Shuttleworth, W. J.: Parameter sensitivity analysis for different complexity land surface models using multicriteria methods, *J. Geophys. Res.*, 111, D20101, <https://doi.org/10.1029/2005JD006377>, 2006.
- Bazilian, M., Rogner, H., Howells, M., Hermann, S., Arent, D., Gielen, D., Steduto, P., Mueller, A., Komor, P., Tol, R. S. J., and Yumkella, K. K.: Considering the energy, water and food nexus: towards an integrated modelling approach, *Energy Policy*, 39, 12, 7896–7906, doi:10.1016/j.enpol.2011.09.039, 2011.

- Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, *Hydrol. Earth Syst. Sci.*, <https://doi.org/10.5194/hess-21-2881-2017>, 2017a.
- 5 Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, *Hydrol. Earth Syst. Sci.*, 21, 589–615, <https://doi.org/10.5194/hess-21-589-2017>, 2017b.
- Becker, M., Meyssignac, B., Xavier, L., Cazenave, A., Alkama, R., and Decharme, B.: Past terrestrial water storage (1980–2008) in the Amazon Basin reconstructed from GRACE and in situ river gauging data, *Hydrol. Earth Syst. Sci.*, 15, 533–546, <https://doi.org/10.5194/hess-15-533-2011>, 2011.
- 10 Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., and Ziese, M.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, *Earth Syst. Sci. Data*, 5, 71–99, <https://doi.org/10.5194/essd-5-71-2013>, 2013.
- Berg, A. A., Famiglietti, J. S., Walker, J. P., and Houser, P. R.: Impact of bias correction to reanalysis products on simulations of North American soil moisture and hydrological fluxes, *J. Geophys. Res.*, 108, 4490, <https://doi.org/10.1029/2002JD003334>, 2003.
- 15 Betts, A. K.: Coupling of water vapor convergence, clouds, precipitation, and land-surface processes, *J Geophys. Res.*, 112, D10108, <https://doi.org/10.1029/2006JD008191>, 2007.
- Beven, K., and Germann, P.: Macropores and water flow in soils revisited, *Water Resources Research*, 49, 3071–3092, <https://doi.org/10.1002/wrcr.20156>, 2013.
- 20 Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W. A., Heinke, J., von Bloh, W. and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, *Water Resour. Res.*, 47, 3, W03509, <https://doi.org/10.1029/2009WR008929>, 2011.
- Bierkens, M. F. P., Global hydrology 2015: State, trends, and directions, *Water Resour. Res.*, 51, 4923–4947, <https://doi.org/10.1002/2015WR017173>, 2015.
- 25 Black, P. E.: Watershed functions, *J. American Water Resour. Assoc.*, 33, 10, 1–11, 1997.
- Boone, A., and Etchevers, P.: An intercomparison of three snow schemes of varying complexity coupled to the same land-surface model: Local scale evaluation at an Alpine site, *J. Hydrometeorol.*, 2, 374–394, 2001
- Boone, A., Masson, V., Meyers, T., and Noilhan, J.: The influence of the inclusion of soil freezing on simulation by a soil-atmosphere-transfer scheme, *J. Appl. Meteorol.*, 39, 1544–1569, [https://doi.org/10.1175/1520-0450\(2000\)039<1544:TIO>2.0.CO;2](https://doi.org/10.1175/1520-0450(2000)039<1544:TIO>2.0.CO;2), 2000.
- 30 Brut, A., Rüdiger, C., Lafont, S., Roujean, J.-L., Calvet, J.-C., Jarlan, L., Gibelin, A.-L., Albergel, C., Le Moigne, P., Soussana, J.-F., Klumpp, K., Guyon, D., Wigneron, J.-P., and Ceschia, E.: Modelling LAI at a regional scale with ISBA-A-gs: comparison with satellite-derived LAI over southwestern France, *Biogeosciences*, 6, 1389–1404, 2009.

- Calvet, J.-C.: Investigating soil and atmospheric plant water stress using physiological and micrometeorological data, *Agr. Forest Meteorol.*, 103, 229–247, 2000.
- Calvet, J.-C., Lafont, S., Cloppet, E., Souverain, F., Badeau, V., and Le Bas, C.: Use of agricultural statistics to verify the interannual variability in land surface models: a case study over France with ISBA-A-gs, *Geosci. Model Dev.*, 5, 37–54, <https://doi.org/10.5194/gmd-5-37-2012>, 2012.
- Calvet, J.-C., Noilhan, J., Roujean, J.-L., Bessemoulin, P., Cabelguenne, M., Olioso, A., and Wigneron, J.-P.: An interactive vegetation SVAT model tested against data from six contrasting sites, *Agr. Forest Meteorol.*, 92, 73–95, 1998.
- Calvet, J.-C., Rivalland, V., Picon-Cochard, C., and Guehl, J.-M.: Modelling forest transpiration and CO₂ fluxes - Response to soil moisture stress, *Agr. Forest Meteorol.*, 124, 143-156, 2004.
- 10 Calvet, J.-C., and Soussana, J.-F.: Modelling CO₂ -enrichment effects using an interactive vegetation SVAT scheme, *Agr. Forest Meteorol.*, 108, 129–152, 2001.
- Campos, A., Ducharne, A., Cheruy, F., Hourdin, F., Polcher, J., and Dupont, J. C.: Response of land surface fluxes and precipitation to different soil bottom hydrological conditions in a general circulation model, *J. Geophys. Res. -Atmos.*, 118, 10, 725–10,739, <https://doi.org/10.1002/jgrd.50627>, 2013.**
- 15 Canal, N., Calvet, J.-C., Decharme, B., Carrer, D., Lafont, S., and Pigeon, G.: Evaluation of root water uptake in the ISBA-A-gs land surface model using agricultural yield statistics over France, *Hydrol. Earth Syst. Sci.*, 18, 4979–4999, <https://doi.org/10.5194/hess-18-4979-2014>, 2014.
- Carrera, M., Bélair, S., and Bilodeau, B.: The Canadian Land Data Assimilation System (CaLDAS): description and synthetic evaluation study, *J. Hydrometeorol.*, 16, 1293–1294, <https://doi.org/10.1175/JHM-D-14-0089.1>, 2015.
- 20 Damerau, K., Anthony, G. P., and van Vliet, O. P. R.: Water saving potentials and possible trade-offs for future food and energy supply, *Glob. Environ. Change*, 39, 15–25, <https://doi.org/10.1016/j.gloenvcha.2016.03.014>, 2016.
- Decharme, B., Alkama, R., Douville, H., Becker, M., and Cazenave, A.: Global evaluation of the ISBA-TRIP continental hydrologic system, Part II: Uncertainties in river routing simulation related to flow velocity and groundwater storage, *J. Hydrometeorol.*, 11, 601–617, <https://doi.org/10.1175/2010JHM1212.1>, 2010.
- 25 Decharme, B., Alkama, R., Papa, F., Faroux, S., Douville, H., and Prigent, C.: Global off-line evaluation of the ISBA-TRIP flood model, *Clim. Dynam.*, 38, 7, 1389–1412, 2012.
- Decharme, B., Boone, A., Delire, C., and Noilhan, J.: Local evaluation of the Interaction between Soil Biosphere Atmosphere soil multilayer diffusion scheme using four pedotransfer functions, *J. Geophys. Res.*, 116, D20126, <https://doi.org/10.1029/2011JD016002>, 2011.
- 30 Decharme, B., and Douville, H.: Introduction of a sub-grid hydrology in the ISBA land surface model, *Clim. Dynam.*, 26, 1, 65–78, <https://doi.org/10.1007/s00382-005-0059-7>, 2006a.
- Decharme, B., and Douville, H.: Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations, *Clim. Dynam.*, 27, 7, 695–713, <https://doi.org/10.1007/s00382-006-0160-6>, 2006b.

- Decharme, B., Martin, E., and Faroux, S.: Reconciling soil thermal and hydrological lower boundary conditions in land surface models, *J. Geophys. Res.*, 118, 1–16, <https://doi.org/10.1002/jgrd.50631>, 2013.
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R.,
5 Fuentes, M., Geer, A. J., Haimberger, L., Healy, S., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., Mc-Nally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Peubey, J., de Rosnay, P., Tavolato, C., Thepaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: Conguration and performance of the data assimilation system, *Q. J. Roy. Meteor. Soc.*, 137, 553–597, <https://doi.org/10.1002/qj.828>, 2011.
- Dewaele, H., Munier, S., Albergel, C., Planque, C., Laanaia, N., Carrer, D., and Calvet, J.-C.: Parameter optimisation for a
10 better representation of drought by LSMs: inverse modelling vs. sequential data assimilation, *Hydrol. Earth Syst. Sci.*, 21, 4861–4878, <https://doi.org/10.5194/hess-21-4861-2017>, 2017.
- Dirmeyer, P. A.: A history and review of the Global Soil Wetness Project (GSWP), *J. Hydrometeorol.*, 12, 729–749, 2011.
- Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel analysis and implications for
our perception of the land surface, *B. Am. Meteorol. Soc.*, 87, 1381–1397, doi:10.1175/BAMS-87-10-1381, 2006.
- 15 Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R. M., and Kidd, R.: Evaluation of the ESA CCI soil moisture product using ground-based observations, *Remote Sens. Environ.*, 162, 380–395, <https://doi.org/10.1016/j.rse.2014.07.023>, 2014.
- Douville H.: Validation and sensitivity of the global hydrologic budget in stand-alone simulations with the ISBA land-surface scheme, *Clim. Dynam.*, 14, 151–171, 1998.
- 20 Draper, C., Mahfouf, J.-F., Calvet, J.-C., Martin, E., and Wagner, W.: Assimilation of ASCAT near-surface soil moisture into the SIM hydrological model over France, *Hydrol. Earth Syst. Sci.*, 15, 3829–3841, <https://doi.org/10.5194/hess-15-3829-2011>, 2011.
- Draper, C., Reichle, R., Lannoy, G. D., and Liu, Q.: Assimilation of passive and active microwave soil moisture retrievals, *Geophys. Res. Lett.*, 39, L04401, <https://doi.org/10.1029/2011GL050655>, 2012.
- 25 Ducharne, A., Golaz, C., Leblois, E., Laval, K., Polcher, J., Ledoux, E., de Marsily, G.: Development of a high resolution runoff routing model, calibration and application to assess runoff from the LMD GCM, *J. Hydrol.*, 280, 207–228, [https://doi.org/10.1016/S0022-1694\(03\)00230-0](https://doi.org/10.1016/S0022-1694(03)00230-0), 2003.
- Entin, J. K., Robock, A., Vinnikov, K. Y., Zabelin, V., Liu, S., Namkhai, A., and Adyasuren, T.: Evaluation of Global Soil Wetness Project soil moisture simulations, *J. Meteorol. Soc. Japan*, 77, 1B, 183–198, 1999.
- 30 Escorihuela, M.J., Chanzy, A., Wigneron, J.P., and Kerr, Y.H.: Effective soil moisture sampling depth of L-band radiometry: a case study, *Remote Sens. Environ.*, 114, 5, 995–1001, <https://doi.org/10.1016/j.rse.2009.12.011>, 2010.
- Fairbairn, D., Barbu, A. L., Mahfouf, J.-F., Calvet, J.-C., and Gelati, E.: Comparing the ensemble and extended Kalman filters for in situ soil moisture assimilation with contrasting conditions, *Hydrol. Earth Syst. Sci.*, 19, 4811–4830, <https://doi.org/10.5194/hess-19-4811-2015>, 2015.

- Faroux, S., Kaptué Tchuenté, A. T., Roujean, J.-L., Masson, V., Martin, E., and Le Moigne, P.: ECOCLIMAP-II/Europe: a twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models, *Geosci. Model Dev.*, 6, 563–582, <https://doi.org/10.5194/gmd-6-563-2013>, 2013.
- 5 Fekete, B. M., Looser, U., Pietroniro, A., and Robarts, R. D.: Rationale for monitoring discharge on the ground, *J. Hydrometeorol.*, 13, 1977–1986, <https://doi.org/10.1175/JHM-D-11-0126.1>, 2012.
- Fuchs, M., Campbell, G. S., Papendick, R. I.: An analysis of sensible and latent heat flow in a partially frozen unsaturated soil, *Soil Sci. Soc. Am. J.*, 42, 379–385, 1978.
- Garrigues, S., Olioso, A., Carrer, D., Decharme, B., Calvet, J.-C., Martin, E., Moulin, S., and Marloie, O.: Impact of climate, vegetation, soil and crop management variables on multi-year ISBA-A-gs simulations of evapotranspiration over a Mediterranean crop site, *Geosci. Model Dev.*, 8, 3033–3053, <https://doi.org/10.5194/gmd-8-3033-2015>, 2015.
- 10 Gibelin, A.-L., Calvet, J.-C., Roujean, J.-L., Jarlan, L., and Los, S. O.: Ability of the land surface model ISBA-A-gs to simulate leaf area index at the global scale: Comparison with satellites products, *J. Geophys. Res.*, 111, D18102, <https://doi.org/10.1029/2005JD006691>, 2006.
- 15 GRDC (Global Runoff Data Centre), Federal Institute of Hydrology, Koblenz, Germany, http://www.bafg.de/GRDC/EN/Home/homepage_node.html, last access: February 2018.
- Goudriaan, J., van Laar, H. H., van Keulen, H., and Louwse, W.: Photosynthesis, CO₂ and plant production, Wheat Growth and Modelling, NATO ASI Series, Plenum Press, New York, Series A, 86, 107–122, 1985.
- Gouttevin, I., Bartsch, A., Krinner, G., and Naeimi, V.: A comparison between remotely-sensed and modelled surface soil moisture (and frozen status) at high latitudes, *Hydrol. Earth Syst. Sci. Discuss.*, <https://doi.org/10.5194/hessd-10-11241-2013>, 2013.
- 20 Grippa, M., Kergoat, L., Frappart, F., Araud, Q., Boone, A., de Rosnay, P., Lemoine, J.-M., Gascoïn, S., Balsamo, G., Otlé, C., Decharme, B., Saux-Picart, S., and Ramillien, G.: Land water storage variability over West Africa estimated by GRACE and land surface models, *Water Resour. Res.*, 47, W05549, <https://doi.org/10.1029/2009WR008856>, 2011.
- 25 Green, W. H., and Ampt, G. A.: Studies on soil physics, 1: The flow of air and water through soils, *J. Agr. Sci.*, 4, 1–24, 1911.
- Guo, Z. and Dirmeyer, P. A. and Hu, Z.-Z. and Gao, X., and Zhao, M.: Evaluation of the Second Global Soil Wetness Project soil moisture simulations: 2. Sensitivity to external meteorological forcing, *J. Geophys. Res. -Atmos.*, 111, D22, <https://doi.org/10.1029/2006JD007845>, 2006.
- 30 Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Parameter estimation of a land surface scheme using multi-criteria methods, *J. Geophys. Res.*, 104, 19,491–19,504, 1999.
- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, 377, 1–2, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.

- Gupta, H. V., Sorooshian, S., Yapo, P. O.: Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information, *Water Resour. Res.* 34, 4, 751–763, 1998.
- Haddeland, I., Heinke, J., Voß, F., Eisner, S., Chen, C., Hagemann, S., and Ludwig, F.: Effects of climate model radiation, humidity and wind estimates on hydrological simulations, *Hydrol. Earth Syst. Sci.*, 16, 305–318, <https://doi.org/10.5194/hess-16-305-2012>, 2012.
- Hagemann, S., Chen, C., Haerter, J.O., Heinke, J., Gerten, D. and Piani, C.: Impact of a statistical bias correction on the projected hydrological changes obtained from three GCMs and two hydrology models, *J. Hydrometeorol.*, 12(4), 556–578, <https://doi.org/10.1175/2011JHM1336.1>, 2011.
- Hanasaki, N., Kanae, S., and Oki, T.: A reservoir operation scheme for global river routing models, *J. Hydrol.*, 327, 1–2, 22–41, <https://doi.org/10.1016/j.jhydrol.2005.11.011>, 2006.
- Hannah, D. M., Demuth, S., Van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, *Hydrol. Proc.*, 25, 1191–1200, <https://doi.org/10.1002/hyp.7794>, 2011.
- Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, *Rev. Geophys.*, 48, RG4004, <https://doi.org/10.1029/2010RG000345>, 2010.
- Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, *Int. J. Clim.*, 34, 3, 623–642, <https://doi.org/10.1002/joc.3711>, 2014.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded dataset of surface temperature and precipitation, *J. Geophys. Res.*, 113, D20119, <https://doi.org/10.1029/2008JD010201>, 2008.
- Hirpa, F., P. Salamon, L. Alfieri, J. Thielen-del Pozo, E. Zsoter, and F. Pappenberger: The effect of reference climatology on global flood forecasting, *J. Hydrometeorol.*, 17, 4, 1131–1145, doi:10.1175/JHM-D-15-0044.1, 2016.
- Huffman, G. J., Adler, R. F., Bolvin, D. T., and Gu, G.: Improving the global precipitation record: GPCP Version 2.1, *Geophys. Res. Lett.*, 36, L17808, <https://doi.org/10.1029/2009GL040000>, 2009.
- IPCC: Climate Change 2014: Synthesis Report, Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Pachauri, R. K., and Meyer, L. A., IPCC, Geneva, Switzerland, 151 pages, 2014.
- Jacobs, C. M. J., Van den Hurk, B. J. J. M., and De Bruin, H. A. R.: Stomatal behaviour and photosynthetic rate of unstressed grapevines in semi-arid conditions, *Agr. Forest Meteorol.*, 80, 111-134, 1996.
- Jägermeyr, J., Gerten, D., and Schaphoff, S., Heinke, J., Lucht, W., and Rockström, J.: Integrated crop water management might sustainably halve the global food gap, *Environ. Res. Lett.*, 11, 2, 025002, <https://doi.org/10.1088/1748-9326/11/2/025002>, 2016.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C.,

- Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, *Bull. Am. Meteorol. Soc.*, **77**, 437–471, [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYP>2.0.CO;2), 1996.
- Kerr, Y., Waldteufel, P., Wigneron, J. P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M., Font, J., Reul, N., Gruhier, C., Juglea, S., Drinkwater, M., Hahne, A., Martin-Neira, M., and Mecklenburg, S.: The SMOS mission: new tool for monitoring key elements of the global water cycle, *Proc. IEEE*, **98**, 666–687, 2010.
- Knutti, R.: Should we believe model predictions of future climate change?, *Phil. Trans. R. Soc. A*, **366**, 4647–4664, <https://doi.org/10.1098/rsta.2008.0169>, 2008.
- Koster, R., Dirmeyer, P., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C. T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C.-H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., and Yamada, T.: Regions of strong coupling between soil moisture and precipitation, *Science*, **305**, 5687, 1138–1140, 2004.
- Krinner, G. N., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Stich, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cy.*, **19**, GB1015, <https://doi.org/10.1029/2003GB002199>, 2005.
- Lafont, S., Zhao, Y., Calvet, J.-C., Peylin, P., Ciais, P., Maignan, F., and Weiss, M.: Modelling LAI, surface water and carbon fluxes at high-resolution over France: comparison of ISBA-A-gs and ORCHIDEE, *Biogeosciences*, **9**, 439–456, <https://doi.org/10.5194/bg-9-439-2012>, 2012.
- Lawford, R., Bogardi, J., Marx, S., Jain, S., Wostl, C. P., Knüppe, K., Ringler, C., Lansigan, F., and Meza, F.: Basin perspectives on the Water-Energy-Food Security Nexus, *Curr. Opin. Environ. Sustain.*, **5**, 6, 607–616, <https://doi.org/10.1016/j.cosust.2013.11.005>, 2013.
- Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J.: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3, *J. Geophys. Res.*, **116**, D19121, <https://doi.org/10.1029/2011JD016187>, 2011.
- Le Moigne, P.: SURFEX scientific documentation, CNRM, Météo-France, Toulouse, France, 237 pp., available at: <http://www.umr-cnrm.fr/surfex/> (last access: February 2018), 2012.
- Li, H., Wigmosta, M. S., Wu, H., Huang, M., Ke, Y., Coleman, A. M., and Leung, L. R.: A physically based runoff routing model for land surface and Earth system models, *J. Hydrometeorol.*, **14**, 3, 808–828, <https://doi.org/10.1175/JHM-D-12-015.1>, 2013.
- Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, *Hydrol. Earth Syst. Sci.*, **15**, 425–436, <https://doi.org/10.5194/hess-15-425-2011>, 2011.
- Liu, J. G., Jia, B. H., Xie, Z. H., and Shi, C. X.: Ensemble simulation of land evapotranspiration in China based on a multi-forcing and multi-model approach, *Adv. Atmos. Sci.*, **33**, 6, 673–684, <https://doi.org/10.1007/s00376-016-5213-0>, 2016.

- Materia, S., Dirmeyer, P. A., Guo, Z., Alessandri, A., and Navarra, A.: The sensitivity of simulated river discharge to land surface representation and meteorological forcings, *J. Hydrometeorol.*, 11, 334–351, 2010.
- Milly, P. C. D., and Shmakin, A. B.: Global modeling of land water and energy balances. Part II: Land-characteristic contributions to spatial variability, *J. Hydrometeorol.*, 3, 3, 301–310, [https://doi.org/10.1175/1525-7541\(2002\)003<0301:GMOLWA>2.0.CO;2](https://doi.org/10.1175/1525-7541(2002)003<0301:GMOLWA>2.0.CO;2), 2002.
- Mätzler, C., and Standley, A.: Relief effects for passive microwave remote sensing, *Int. J. Remote Sens.*, 21, 12, 2403–2412, <https://doi.org/10.1080/01431160050030538>, 2000.
- Muerth, M. J., Gauvin St-Denis, B., Ricard, S., Velázquez, J. A., Schmid, J., Minville, M., Caya, D., Chaumont, D., Ludwig, R., and Turcotte, R.: On the need for bias correction in regional climate scenarios to assess climate change impacts on river runoff, *Hydrol. Earth Syst. Sci.*, 17, 1189–1204, <https://doi.org/10.5194/hess-17-1189-2013>, 2013.
- Muñoz-Sabater, J.: Incorporation of Passive Microwave Brightness Temperatures in the ECMWF Soil Moisture Analysis, *Remote Sens.*, 7, 5758–5784, <https://doi.org/10.3390/rs70505758>, 2015.
- Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through. Part I. A conceptual models discussion of principles, *J. Hydrol.*, 10, 282–290, 1970.
- Nasonova, O. N., Gusev, Y. M., and Kovalev, Y. E.: Impact of uncertainties in meteorological forcing data and land surface parameters on global estimates of terrestrial water balance components, *Hydrol. Processes*, 25, 7, 1074–1090, <https://doi.org/10.1002/hyp.7651>, 2011.
- Ngo-Duc, T., Polcher, J., and Laval, K.: A 53-year forcing data set for land surface models, *J. Geophys. Res.*, 110, D6, <https://doi.org/10.1029/2004JD005434>, 2005.
- Noilhan, J., and Mahfouf, J.-F.: The ISBA land surface parameterization scheme, *Global Planet. Change*, 13, 145–149, 1996.
- Noilhan, J., and Planton, S.: A simple parameterisation of Land Surface Processes for meteorological model, *Mon. Weather Rev.*, 117, 356–549, 1989.
- Overgaard, J., Rosbjerg, D., and Butts, M. B.: Land-surface modelling in hydrological perspective – a review, *Biogeosciences*, 3, 229–241, <https://doi.org/10.5194/bg-3-229-2006>, 2006.
- Papadimitriou, L. V., Koutroulis, A. G., Grillakis, M. G., and Tsanis, I. K.: The effect of GCM biases on global runoff simulations of a land surface model, *Hydrol. Earth Syst. Sci.*, 21, 4379–4401, <https://doi.org/10.5194/hess-21-4379-2017>, 2017.
- Pappenberger, F., Cloke, H. L., Balsamo, G., Ngo-Duc, T., and Oki, T.: Global runoff routing with the hydrological component of the ECMWF NWP system, *Int. J. Clim.*, 30, 14, 2155–2174, <https://doi.org/10.1002/joc.2028>, 2010.
- Pappenberger, F., Dutra, E., Wetterhall, F., and Cloke, H. L.: Deriving global flood hazard maps of fluvial floods through a physical model cascade, *Hydrol. Earth Syst. Sci.*, 16, 4143–4156, <https://doi.org/10.5194/hess-16-4143-2012>, 2012.
- Parrens, M., Zakharova, E., Lafont, S., Calvet, J.-C., Kerr, Y., Wagner, W., and Wigneron, J.-P.: Comparing soil moisture retrievals from SMOS and ASCAT over France, *Hydrol. Earth Syst. Sci.*, 16, 2, 423–440, <https://doi.org/10.5194/hess-16-423-2012>, 2012.

- Parrens, M., Calvet, J.-C., de Rosnay, P., and Decharme, B.: Benchmarking of L-band soil microwave emission models, *Remote Sens. Environ.*, 140, 407–419, <https://doi.org/10.1016/j.rse.2013.09.017>, 2014.
- PGF (Global Meteorological Forcing Dataset for land surface modeling), Terrestrial Hydrology Research Group, Princeton University, Princeton, NJ, USA, <http://hydrology.princeton.edu/data.pgf.php>, last access: February 2018.
- 5 Planton, S., Lionello, P., Artale, V., Aznar, R., Carillo, A., Colin, J., Congedi, L., Dubois, C., Elizalde Arellano, A., Gualdi, S., Hertig, E., Jorda Sanchez, G., Li, L., Jucundus, J., Piani, C., Ruti, P., Sanchez-Gomez, E., Sannino, G., Sevault, F., and Somot, S.: The climate of the Mediterranean region in future climate projections, in: *The Climate of the Mediterranean Region*, Chapter 8, 1st Edition, edited by Lionello, P., Elsevier, 2012.
- Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H. Kanae, S., Oki, T., Incorporating anthropogenic water regulation modules into a land surface model, *J. Hydrometeorol.*, 13, 1, 255–269, doi:10.1175/JHM-D-11-013.1, 2012.
- 10 Polcher, J., Piles, M., Gelati, E., Barella-Ortiz, A., and Tello, M.: Comparing surface-soil moisture from the SMOS mission and the ORCHIDEE land-surface model over the Iberian Peninsula, *Remote Sens. Environ.*, 174, 69–81, <https://doi.org/10.1016/j.rse.2015.12.004>, 2016.
- Oki, T., and Sud, Y. C.: Design of Total Runoff Integrating Pathways (TRIP) – a global river channel network, *Earth Interact.*, 2, 1–37, 1998.
- 15 Reichle, R., Crow, W., and Keppenne, C.: An adaptive Ensemble Kalman Filter for soil moisture data assimilation, *Water Resour. Res.*, 44, W03243, <https://doi.org/10.1029/2007WR006357>, 2008.
- Ringler, C., Bhaduri, A. and Lawford, R.: The nexus across water, energy, land and food (WELF): Potential for improved resource use efficiency?, *Curr. Opin. Environ. Sustain.*, 5, 6, 617–624, <https://doi.org/10.1016/j.cosust.2013.11.002>, 2013.
- 20 Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, *Bull. Am. Meteorol. Soc.*, 85, 3, 381–394, <https://doi.org/10.1175/BAMS-85-3-381>, 2004.
- Rohde, R., Muller, R. A., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtele, J., Groom, D., and Wickham, C.:
25 **A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011, *Geoinformatics & Geostatistics: An Overview*, <https://doi.org/10.4172/2327-4581.1000101>, 2013.**
- Rost, S., Gerten, D., Hoff, H., Lucht, W., Falkenmark, M., and Rockström, J.: Global potential to increase crop production through water management in rainfed agriculture, *Environ. Res. Lett.*, 4, 4, 044002, <https://doi.org/10.1088/1748-9326/4/4/044002>, 2009.
- 30 Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: the earth2Observe Tier-1 dataset, *Earth Syst. Sci. Data*, 9, 389–413, <https://doi.org/10.5194/essd-9-389-2017>, 2017.

- Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., and Ziese, M.: GPCP Full Data Reanalysis Version 6.0 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data, Global Precipitation Climatology Centre (GPCC) at Deutscher Wetterdienst, https://doi.org/10.5676/DWD_GPCP/FD_M_V6_050, 2011.
- 5 Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCP's new land surface precipitation climatology based on quality-controlled insitu data and its role in quantifying the global water cycle, *Theor. Appl. Climatol.*, 115, 15–40, <https://doi.org/10.1007/s00704-013-0860-x>, 2014.
- Sheffield, J., Goteti, G., and Wood, E. F.: Development of a 50-Year high-resolution global dataset of meteorological forcings for land surface modeling, *J. Climate*, 19, 3088–3111, <https://doi.org/10.1175/JCLI3790.1>, 2006.
- 10 Sheffield, J., Ziegler, A. D., Wood, E. F., and Chen, Y.: Correction of the high-latitude rain day anomaly in the NCEP-NCAR reanalysis for land surface hydrological modeling. *J. Climate*, 17, 3814–3828, 2004.
- Sippel, S., Otto, F. E. L., Forkel, M., Allen, M. R., Guillod, B. P., Heimann, M., Reichstein, M., Seneviratne, S. I., Thonicke, K., and Mahecha, M. D.: A novel bias correction methodology for climate impact simulations, *Earth Syst. Dynam.*, 7, 71–88, <https://doi.org/10.5194/esd-7-71-2016>, 2016.
- 15 Smith, P. C., Ciais, P., Peylin, P., De Noblet-Ducoudré, N., Viovy, N., Meurdesoif, Y., and Bondeau, A.: European-wide simulations of croplands using an improved terrestrial biosphere model: 2. interannual yields and anomalous CO₂ fluxes in 2003, *J. Geophys. Res.*, 115, G04028, <https://doi.org/10.1029/2009JG001041>, 2010a.
- Smith, P. C., De Noblet-Ducoudré, N., Ciais, P., Peylin, P., Viovy, N., Meurdesoif, Y., and Bondeau, A.: European-wide simulations of croplands using an improved terrestrial biosphere model: phenology and productivity, *J. Geophys. Res.*, 20 115, G01014, <https://doi.org/10.1029/2008JG000800>, 2010b.
- Stoffelen, A., Aaboe, S., Calvet, J.-C., Cotton, J., De Chiara, G., Figua-Saldana, J., Mouche, A. A., Portabella, M., Scipal, K., Wagner, W.: Scientific developments and the EPS-SG scatterometer, *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, 10, 2086–2097, <https://doi.org/10.1109/JSTARS.2017.2696424>, 2017.
- Swenson, S., Wahr, J., and Milly, P. C. D.: Estimated accuracies of regional water storage variations inferred from the Gravity Recovery and Climate Experiment (GRACE), *Water Resour. Res.*, 39, 1223, <https://doi.org/10.1029/2002WR001808>, 2003.
- Szczypta, C., Calvet, J.-C., Maignan, F., Dorigo, W., Baret, F., and Ciais, P.: Suitability of modelled and remotely sensed essential climate variables for monitoring Euro-Mediterranean droughts, *Geosci. Model Dev.*, 7, 931–946, <https://doi.org/10.5194/gmd-7-931-2014>, 2014.
- 30 Szczypta, C., Decharme, B., Carrer, D., Calvet, J.-C., Lafont, S., Somot, S., Faroux, S., and Martin, E.: Impact of precipitation and land biophysical variables on the simulated discharge of European and Mediterranean rivers, *Hydrol. Earth Syst. Sci.*, 16, 3351–3370, <https://doi.org/10.5194/hess-16-3351-2012>, 2012.
- van Beek, L. P. H., Eikelboom, T., van Vliet, M. T. H., and Bierkens, M. F. P.: A physically based model of global freshwater surface temperature, *Water Resour. Res.*, 48, W09530, <https://doi.org/10.1029/2012WR011819>, 2012.

- van der Schrier, G., van den Besselaar, E. J. M., Klein Tank, A. M. G., and Verver, G.: Monitoring European average temperature based on the E-OBS gridded data set, *J. Geophys. Res. -Atmos.*, 118, 5120–5135, <https://doi.org/10.1002/jgrd.50444>, 2013.
- van Vliet, M. T. H., Yearsley, J. R., Ludwig, F., Vögele, S., Lettenmaier, D.P., Kabat, P.: Vulnerability of US and European electricity supply to climate change, *Nature Climate Change*, 2, 9, 676–681, <https://doi.org/10.1038/nclimate1546>, 2012.
- 5 Vergnes, J.-P. and Decharme, B.: A simple groundwater scheme in the TRIP river routing model: global off-line evaluation against GRACE terrestrial water storage estimates and observed river discharges, *Hydrol. Earth Syst. Sci.*, 16, 3889–3908, <https://doi.org/10.5194/hess-16-3889-2012>, 2012.
- Vergnes, J.-P., Decharme, B. and Habets, F., Introduction of groundwater capillary rises using subgrid spatial variability of topography into the ISBA land surface model, *J. Geophys. Res. -Atmos.*, 119, 19, 11065–11086, <https://doi.org/10.1002/2014JD021573>, 2014.
- 10 Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, *Geography Compass*, 1, 4, 901–931, 2007.
- Wagner, W., Lemoine, G., Borgeaud, M., and Rott, H.: A study of vegetation cover effects on ERS scatterometer data, *IEEE Trans. Geosci. Remote Sens.*, 37, 938–948, 1999.
- 15 Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, *Water Resour. Res.*, 50, 7505–7514, <https://doi.org/10.1002/2014WR015638>, 2014.
- Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O. and Best, M.: Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century, *J. Hydrometeorol.*, 12, 5, 823–848, <https://doi.org/10.1175/2011JHM1369.1>, 2011.
- 20 Widén-Nilsson, E., Halldin, S., and Xu, C.: Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation, *J. Hydrol.*, 340, 1–2, 105–118, <https://doi.org/10.1016/j.jhydrol.2007.04.002>, 2007.
- 25 Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee III, R. B., Smith, G. L., and Cooper, J. E.: Clouds and the Earth's radiant energy system (CERES): an Earth observing system experiment, *Bull. Am. Meteorol. Soc.*, 77, 5, 853–868, [https://doi.org/10.1175/1520-0477\(1996\)077<0853:CATERE>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2), 1996.
- Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N. V., Wang, Y., Privette, J. L., Huemmrich, K. F., Fensholt, R., Sandholt, I., Weiss, M., Ahl, D. E., Gower, S. T., Nemani, R. R., Knyazikhin, Y., and Myneni, R. B.: MODIS leaf area index products: From validation to algorithm improvement, *IEEE Trans. Geosci. Remote Sens.*, 44, 1885–1898, <https://doi.org/10.1109/TGRS.2006.871215>, 2006.
- 30 Yearsley, J. R.: A grid-based approach for simulating stream temperature, *Water Resour. Res.*, 48, W03506, <https://doi.org/10.1029/2011WR011515>, 2012.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, 44, W09417, <https://doi.org/10.1029/2007WR006716>, 2008.

Yoshimura, K., and Kanamitsu, M.: Incremental correction for the dynamical downscaling of ensemble mean atmospheric fields, *Mon. Weather Rev.*, 141, 3087–3101, <https://doi.org/10.1175/MWR-D-12-00271.1>, 2013.

5 Zhang, T., Stackhouse, P. W., Gupta, S. K., Cox, S. J., and Mikovitz, J. C.: The validation of the GEWEX SRB surface longwave flux data products using BSRN measurements, *J. Quant. Spectrosc. Radiat. Transfer*, 150, 134–147, <https://doi.org/10.1016/j.jqsrt.2014.07.013>, 2015.

Zhang, T., Stackhouse, P. W., Gupta, S. K., Cox, S. J., and Mikovitz, J. C., and Hinkelman, L. M.: The validation of the GEWEX SRB surface shortwave flux data products using BSRN measurements: A systematic quality control, production and application approach, *J. Quant. Spectrosc. Radiat. Transfer*, 122, 127–140, <https://doi.org/10.1016/j.jqsrt.2012.10.004>, 2013.

Zhou, T., Nijssen, B., Gao, H., and Lettenmaier, D.: The contribution of reservoirs to global land surface water storage variations, *J Hydrometeorol.*, 17, 1, 309–325, <https://doi.org/10.1175/JHM-D-15-0002.1>, 2016.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R. R., and Myneni, R. B.: Global data sets of vegetation leaf area index (LAI)3g and fraction of photosynthetically active radiation (FPAR)3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3g) for the period 1981 to 2011, *Remote Sens.*, 5, 927–948, 2013.

Table 1. River discharge scores (RE_{μ} , RE_{σ} , MC) computed at downstream GRDC gauges. For each gauge and score pair, the best/worst simulation is highlighted in bold black/grey.

	RE_{μ}				RE_{σ}				MC			
	ERA-I	P-ERA	WFDEI	PGF	ERA-I	P-ERA	WFDEI	PGF	ERA-I	P-ERA	WFDEI	PGF
Volga	0.08	0.46	0.11	0.19	1.19	1.93	1.62	1.17	0.14	0.18	0.21	0.16
Danube	0.03	0.19	0.05	0.04	0.52	0.88	0.41	0.27	0.77	0.73	0.86	0.87
Dnepr	0.69	1.45	0.53	0.79	2.42	4.39	2.61	2.65	0.33	0.35	0.32	0.33
Don	1.24	3.24	1.14	0.99	5.75	10.81	6.48	4.99	0.30	0.33	0.35	0.40
Severnaya Dvina	-0.11	0.05	-0.18	0.03	-0.47	-0.37	-0.40	-0.29	0.03	0.10	0.15	-0.09
Neva	-0.05	0.32	-0.09	0.13	1.49	2.47	1.54	1.61	-0.41	-0.42	-0.38	-0.32
Vistula	-0.05	0.00	-0.24	-0.12	0.75	1.00	0.39	0.36	0.64	0.62	0.65	0.69
Ural	1.99	2.91	1.44	0.52	4.24	5.94	3.69	1.81	0.27	0.29	0.33	0.36
Rhine	0.01	0.24	0.19	0.09	0.40	0.81	0.43	0.34	0.79	0.82	0.54	0.61
Elbe	-0.03	0.55	-0.08	-0.05	0.48	1.50	0.35	0.26	0.76	0.77	0.81	0.76
Oder	-0.21	0.10	-0.36	-0.27	0.71	1.45	0.46	0.44	0.62	0.63	0.60	0.60
Rhone	-0.05	0.12	0.02	0.00	0.43	0.66	0.26	0.30	0.82	0.85	0.76	0.76
Ebro	-0.21	0.28	0.04	0.09	-0.03	0.58	0.44	0.40	0.78	0.84	0.79	0.78
Nemunas	-0.10	0.31	-0.23	0.01	0.75	1.67	0.76	0.77	0.65	0.73	0.69	0.81
Po	-0.36	-0.17	-0.10	-0.13	-0.14	0.12	-0.03	0.02	0.71	0.78	0.74	0.74
Tejo	-0.09	0.31	-0.22	-0.26	0.07	0.51	-0.12	-0.20	0.87	0.90	0.86	0.84
Daugava	-0.02	0.44	-0.08	0.11	-0.07	0.41	0.04	-0.06	0.55	0.62	0.56	0.64
Duero	0.00	0.61	0.12	0.12	0.14	0.82	0.31	0.23	0.87	0.90	0.88	0.85
Vuoksi	0.10	0.50	-0.05	-0.03	1.92	2.85	1.68	1.70	0.13	0.14	0.15	0.15
Guadiana	0.64	1.58	0.65	0.58	0.18	0.74	0.15	0.03	0.78	0.83	0.79	0.76
Mezen	-0.21	-0.07	-0.25	-0.15	-0.50	-0.40	-0.41	-0.33	0.31	0.42	0.44	0.10
Narva	-0.09	0.30	-0.15	0.00	1.36	2.48	1.65	1.39	0.29	0.37	0.31	0.34
Kemijoki	0.00	0.19	-0.24	-0.17	-0.12	0.09	-0.21	-0.26	0.58	0.60	0.54	0.31
Guadalquivir	0.47	1.73	0.60	0.52	0.25	1.22	0.38	0.23	0.80	0.86	0.87	0.82
Goeta	-0.06	0.06	0.02	0.02	0.49	0.77	0.74	0.60	0.64	0.63	0.61	0.56
Glama	0.08	0.10	-0.02	-0.21	0.42	0.34	0.35	-0.14	0.62	0.63	0.77	0.47
Weser	-0.31	0.25	-0.28	-0.22	-0.03	0.79	0.02	-0.02	0.85	0.89	0.87	0.82
Kymijoki	0.20	0.55	-0.01	0.01	1.63	2.48	1.49	1.31	0.33	0.30	0.32	0.32
Torneaelven	-0.01	-0.04	-0.14	-0.22	-0.03	-0.01	-0.05	-0.28	0.69	0.63	0.83	0.68

Meuse	-0.08	0.15	0.16	0.26	-0.01	0.27	0.23	0.16	0.85	0.89	0.91	0.90
Kem	-0.14	0.05	-0.25	-0.10	0.29	0.50	0.22	0.38	0.06	0.19	0.16	0.14
Kokemaenjoki	0.08	0.46	-0.15	-0.25	0.67	1.25	0.52	0.31	0.55	0.54	0.49	0.47
Kovda	-0.21	0.04	0.22	0.10	1.54	2.44	2.90	2.54	0.30	0.27	0.28	0.29
Indalsaelven	-0.22	-0.19	-0.19	-0.28	0.26	0.34	0.47	0.32	0.50	0.48	0.58	0.31
Luleaelven	-0.13	-0.12	-0.13	-0.28	2.40	2.42	2.40	1.62	0.33	0.35	0.26	0.33

Table 2. River discharge scores (AC, KGE) computed at downstream GRDC gauges. For each gauge and score pair, the best/worst simulation is highlighted in bold black/grey.

	AC				KGE			
	ERA-I	P-ERA	WFDEI	PGF	ERA-I	P-ERA	WFDEI	PGF
Volga	0.34	0.32	0.30	0.31	-0.47	-1.14	-0.80	-0.45
Danube	0.83	0.85	0.84	0.86	0.43	0.06	0.56	0.70
Dnepr	0.57	0.57	0.53	0.57	-1.61	-3.67	-1.75	-1.84
Don	0.29	0.32	0.30	0.41	-4.92	-10.30	-5.62	-4.12
Severnaya Dvina	0.46	0.44	0.43	0.35	-0.09	0.02	0.04	-0.13
Neva	0.23	0.24	0.16	0.17	-1.05	-1.86	-1.07	-1.09
Vistula	0.70	0.68	0.72	0.71	0.17	-0.07	0.42	0.51
Ural	0.21	0.25	0.26	0.23	-3.74	-5.65	-3.02	-0.99
Rhine	0.85	0.90	0.78	0.79	0.55	0.13	0.34	0.48
Elbe	0.71	0.77	0.80	0.72	0.46	-0.62	0.59	0.65
Oder	0.62	0.64	0.66	0.66	0.16	-0.50	0.29	0.35
Rhone	0.85	0.89	0.80	0.82	0.53	0.32	0.65	0.61
Ebro	0.69	0.77	0.71	0.70	0.69	0.34	0.51	0.53
Nemunas	0.63	0.66	0.62	0.72	0.17	-0.72	0.15	0.21
Po	0.82	0.87	0.79	0.80	0.52	0.70	0.72	0.71
Tejo	0.76	0.82	0.81	0.83	0.83	0.40	0.72	0.63
Daugava	0.63	0.64	0.59	0.63	0.54	0.29	0.55	0.62
Duero	0.75	0.77	0.67	0.70	0.81	-0.03	0.64	0.70
Vuoksi	0.32	0.33	0.37	0.35	-1.11	-2.02	-0.89	-0.90
Guadiana	0.66	0.72	0.75	0.75	0.30	-0.75	0.30	0.37
Mezen	0.52	0.53	0.48	0.43	0.12	0.29	0.26	0.03
Narva	0.48	0.55	0.57	0.52	-0.53	-1.57	-0.79	-0.54
Kemijoki	0.47	0.47	0.47	0.39	0.57	0.55	0.44	0.24
Guadalquivir	0.69	0.71	0.72	0.72	0.43	-1.12	0.28	0.41
Goeta	0.52	0.53	0.50	0.47	0.39	0.14	0.16	0.25
Glama	0.60	0.62	0.63	0.54	0.43	0.49	0.58	0.41
Weser	0.82	0.86	0.84	0.76	0.65	0.16	0.69	0.71
Kymijoki	0.53	0.49	0.51	0.50	-0.78	-1.63	-0.64	-0.48
Torneaelven	0.53	0.55	0.52	0.48	0.69	0.62	0.77	0.52

Meuse	0.83	0.89	0.90	0.87	0.83	0.67	0.70	0.68
Kem	0.52	0.53	0.45	0.55	0.01	0.05	0.10	0.05
Kokemaenjoki	0.62	0.61	0.56	0.57	0.19	-0.41	0.26	0.34
Kovda	0.13	0.12	0.08	0.17	-0.70	-1.55	-2.00	-1.64
Indalsaelven	0.46	0.46	0.41	0.40	0.39	0.35	0.34	0.19
Luleaelven	0.20	0.22	0.24	0.23	-1.49	-1.51	-1.51	-0.78

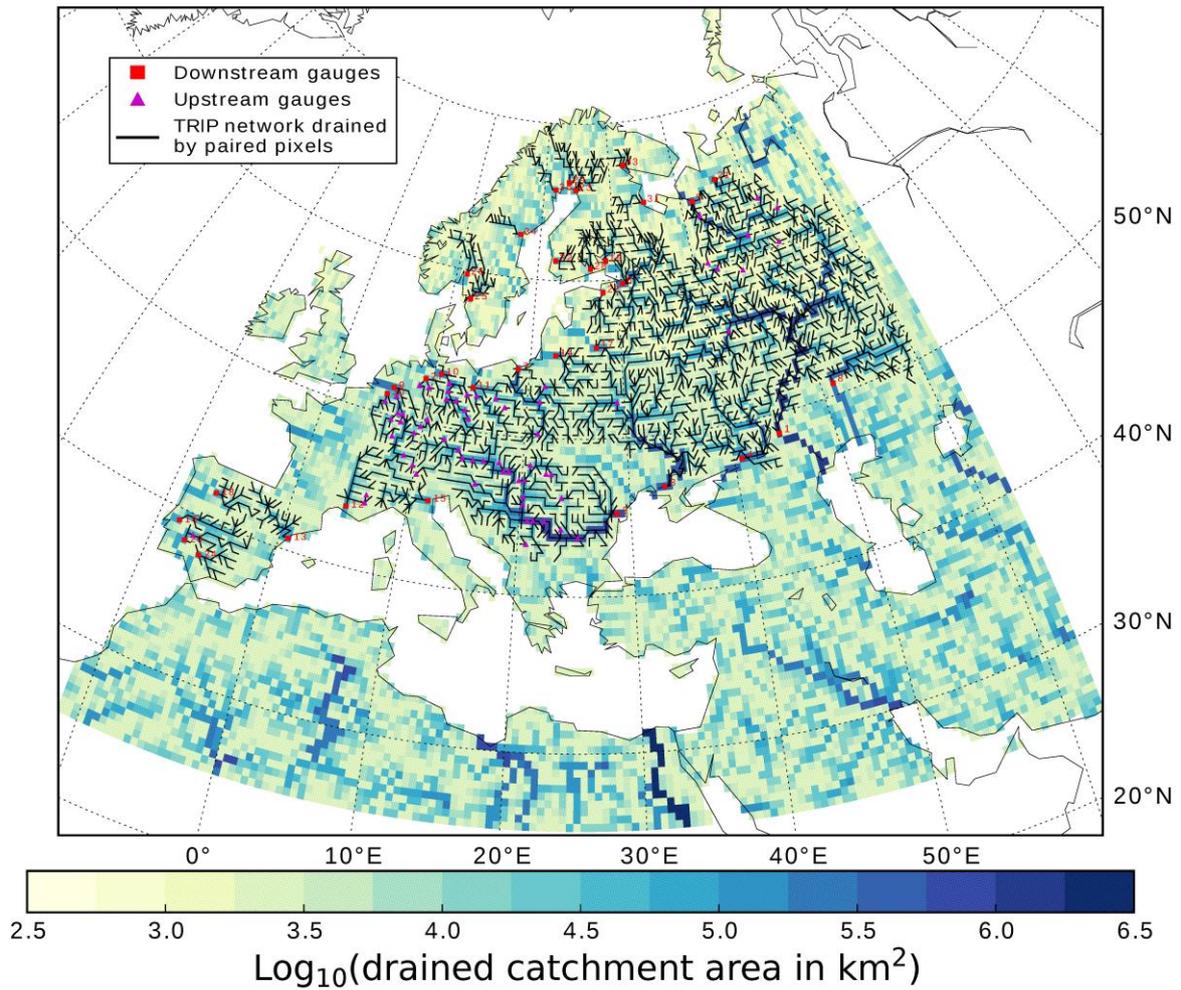


Figure 1: Selected GRDC discharge gauges and corresponding drained CTRIP river network. The colour map categorises reflect the logarithm of the upstream catchment area of each pixel. The numbers associated to downstream gauges are sorted by descending upstream area and refer to the index N in Fig. 2. For readability, gauges are mapped on the paired pixel.

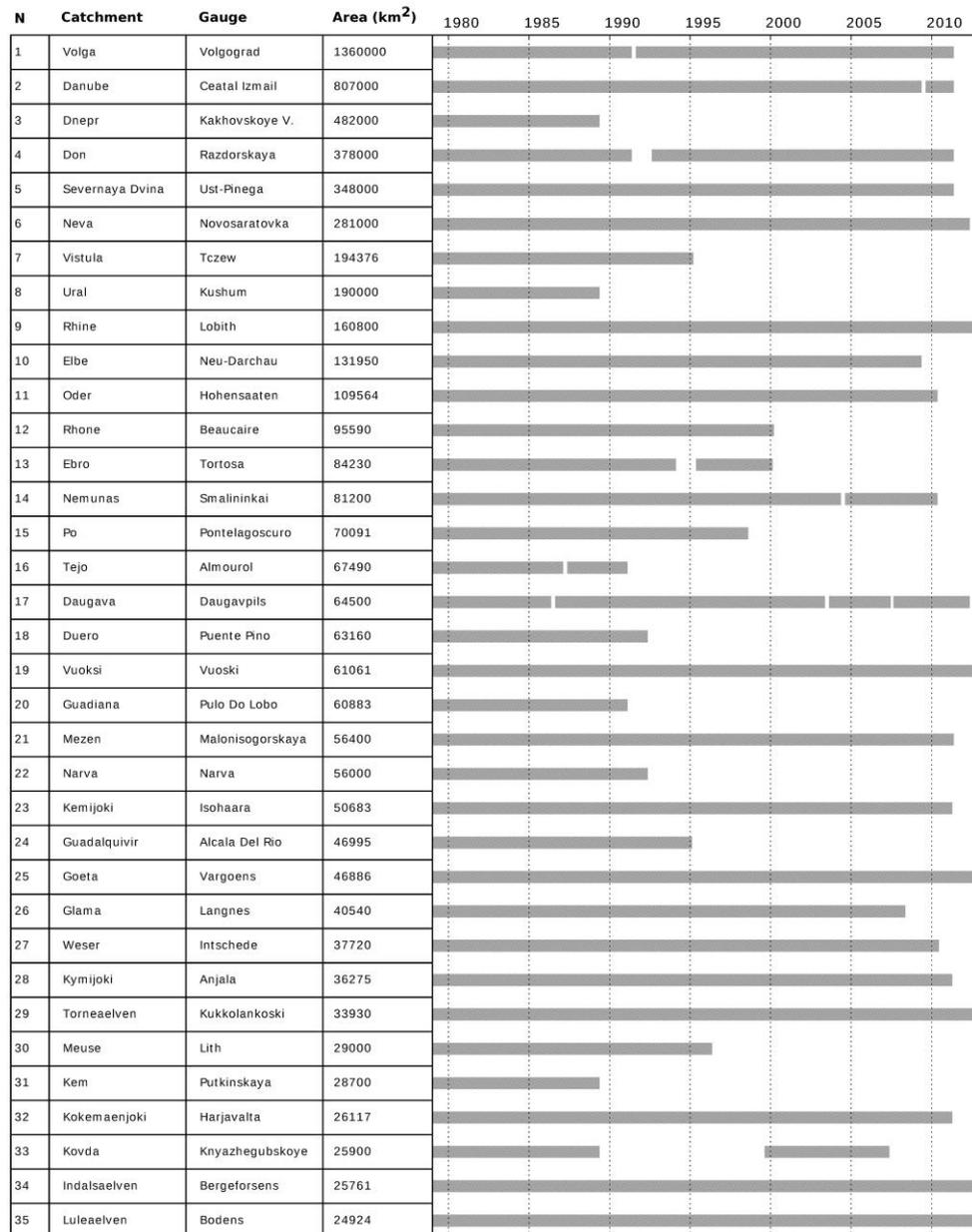


Figure 2: Discharge record availability and drained areas of downstream GRDC gauges. The indices N are mapped in Fig. 1.

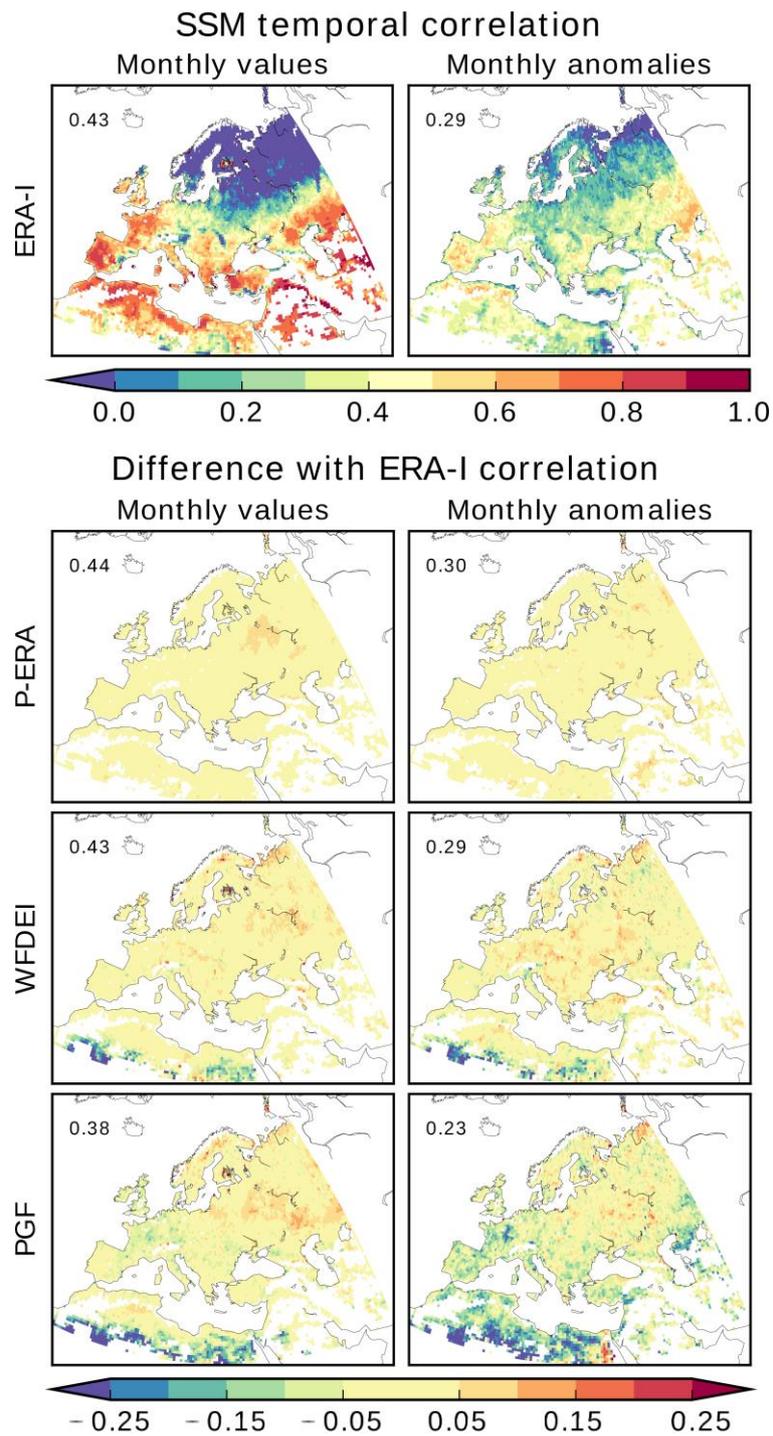


Figure 3: Temporal correlation between modelled and remotely sensed ESA-CCI SSM for monthly averages and anomalies. Global correlation coefficients are reported in the upper left corner of each map.

SSM correlations by month

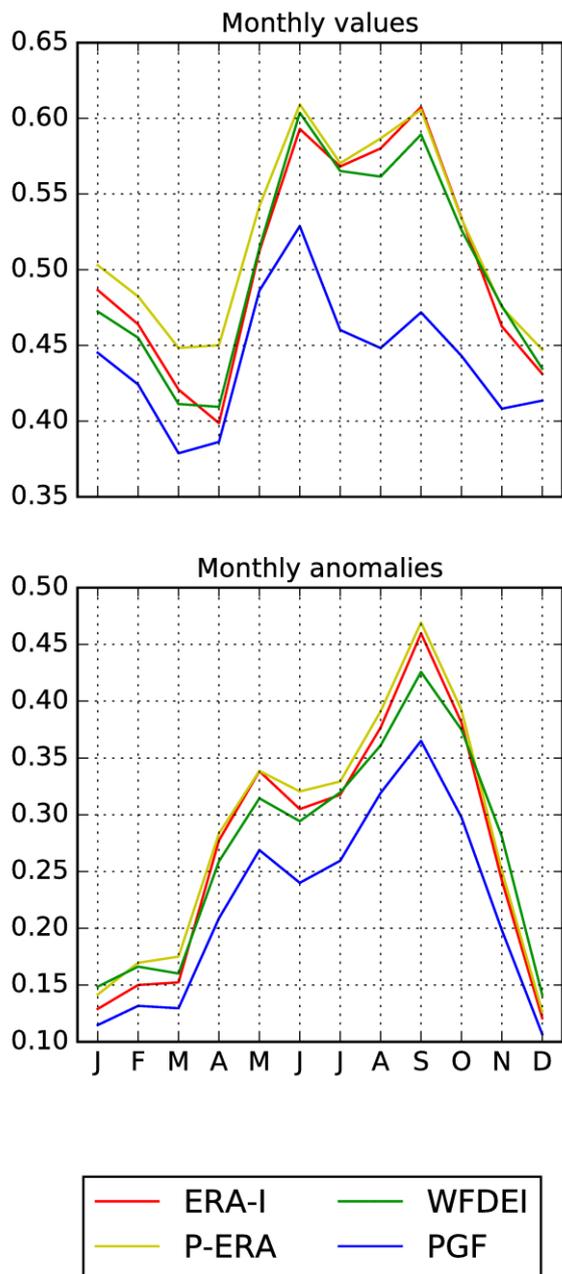


Figure 4: Annual cycles of correlation between modelled and remotely sensed ESA-CCI SSM for monthly averages and anomalies. Coefficients are computed separately for each calendar month.

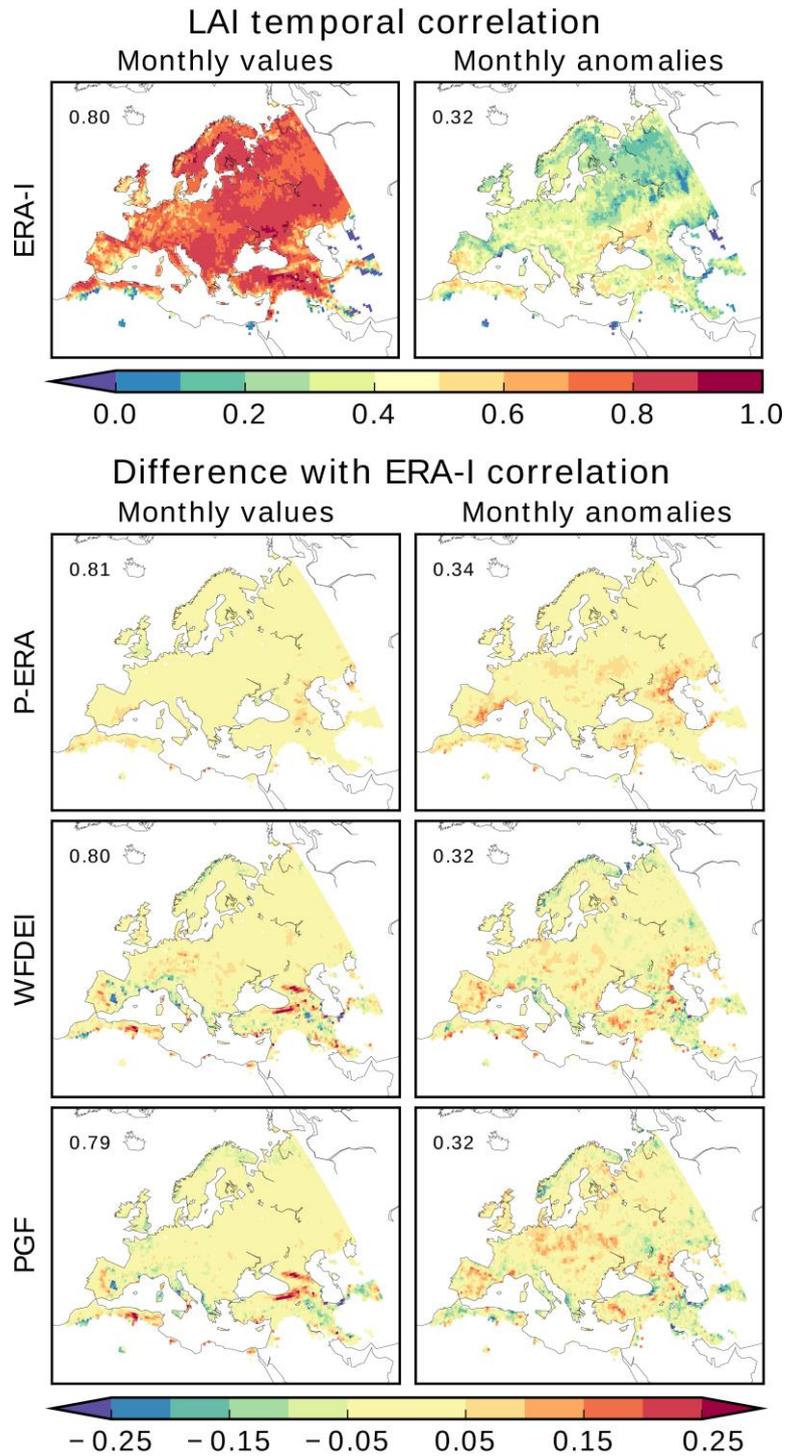


Figure 5: Temporal correlation between modelled and remotely sensed GIMMS LAI for monthly averages and anomalies. Global correlation coefficients are reported in the upper left corner of each map.

LAI correlations by month

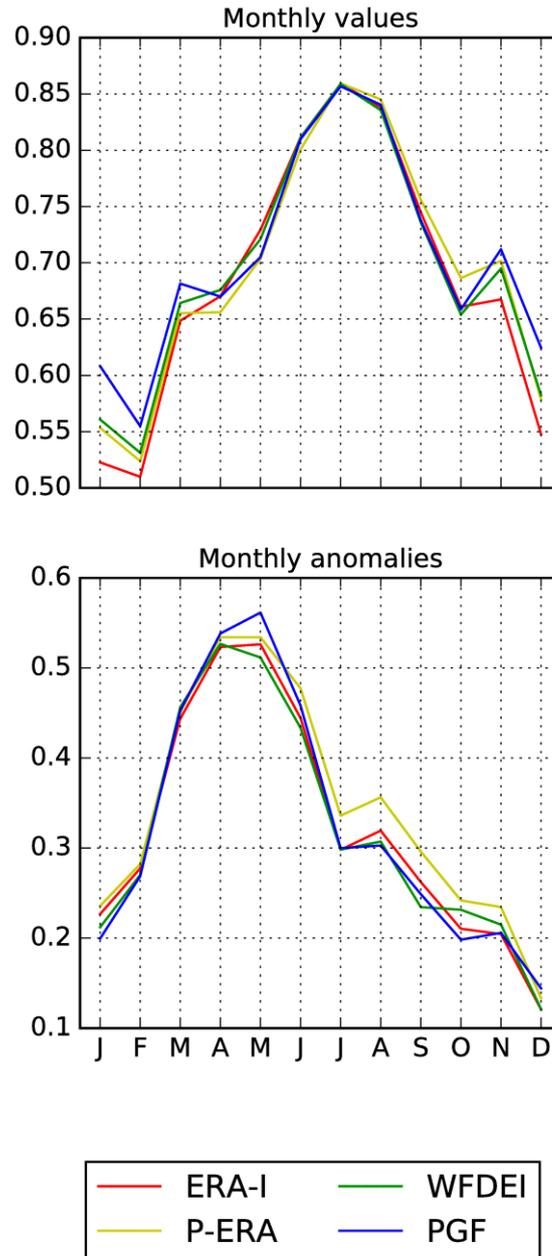


Figure 6: Annual cycles of correlation between modelled and remotely sensed GIMMS LAI for monthly averages and anomalies. Coefficients are computed separately for each calendar month.

Discharge scores (all gauges)

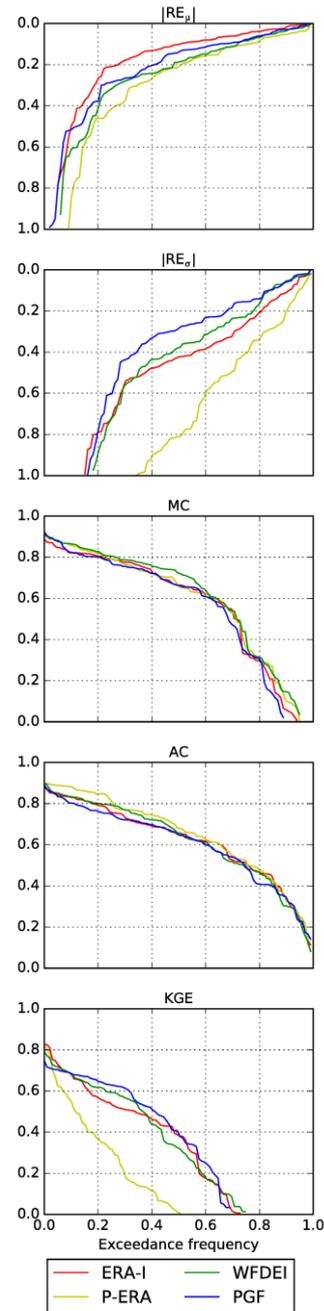


Figure 7: Cumulated frequency plots of river discharge scores computed at all GRDC gauges. $|RE_{\mu}|$ and $|RE_{\sigma}|$ values larger than 1 are not shown. MC, AC and KGE are truncated at 0.